Real-time E-Commerce Comment Classification Using Big Data Processing

Binh-Hau Tran[*,†], Trong-Hop Do[*,†]

[*] University of Information Technology, Ho Chi Minh City, Vietnam.
[†] Vietnam National University, Ho Chi Minh City, Vietnam.

***Abstract*—The 4.0 industrial revolution, and recently the digital transformation trend, have pushed organizations and businesses to apply information technology in their daily activities. Business owners, operators, and managers are increasingly interested in exploiting information from software data and social networks to increase the competitiveness of their units. Collecting and analyzing comments is a method that applies natural language processing (NLP) techniques, combining machine learning, deep learning, and big data to Provides truly valuable information. This article analyzes comments on e-commerce websites, applying the above techniques, thereby comparing and evaluating the advantages and limitations of the methods. From there, the problem can be expanded for many different purposes, as long as it is related to the problem of linguistic analysis to provide recommended information.**

***Index Terms*—Natural Language Processing, Big Data, Deep Learning, Text classification**

I. Introduction

Although social networks or e-commerce channels have shown the connection between customers and potential customers, there are still many subjective comments and assessments of reviewers, sometimes causing confusion. Shoppers become confused by the amount of multi-dimensional information to make decisions. On the other hand, companies providing products and services also want to collect these reviews to improve product quality. What if the review data is too much, but the information is chaotic, making that brand's strategic planners unable to make decisions. In this topic, with the goal of building a miniature model of how to collect, analyze and display comment classification results on some e-commerce sites for certain brands based on applications. big data processing technology. Within the scope of the project, we choose the Tiki e-commerce system, with the item being a laptop to carry out the project. The research goal of this thesis is to build a comment classification system on e-commerce sites applying real-time big data processing technology. Steps to build the system:

- Data collection: data is collected online through connection to the e-commerce system.
- Data preprocessing: checking, standardizing data, removing less valuable information in the data.
- Building a classification model: using big data processing techniques combined with natural language processing to make appropriate judgments.
- Data display: classification results are displayed as graphs, on the web platform.
- Update data: data is updated automatically when the system is operated.
- Analyze and categorize: based on comment types along with the number of comments so that stakeholders, including customers and brand owners, can take actions consistent with their strategies.

[1] [2]

II. Methodology

A. Prepare data for machine learning problems

In problems applying machine learning and deep learning techniques, the contribution of data is extremely large, determining the effectiveness of the model and prediction results. However, creating a useful dataset requires a lot of effort. If a problem is built from scratch, without previously having a standard data set, building this data set consumes up to 80% of the resources of the entire problem. In general, building a dataset usually goes through the following steps: Data Collection, Data Cleansing, Data Transformation and Data Separation.

B. Overview of Big Data

Big data is a term that refers to large and complex data sets that are difficult to process using traditional database management tools. These data sets can be structured, unstructured, or semi-structured. The challenge of this data set is to track, monitor, store, search, share, transform, analyze and abstract. Big data is applied in many fields such as finance, healthcare, e-commerce, digital marketing, retail, and many other fields. Software that implements big data is present in our daily lives, such as google search, facebook, youtube, tiktok, instagram, ebanking, grab, tiki, agoda, airbnb, booking,... [3]

979-8-3503-3094-6/24/$31.00 ©2024 IEEE

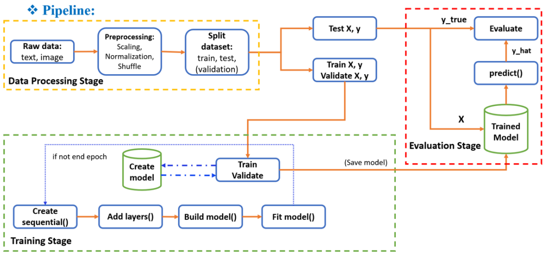
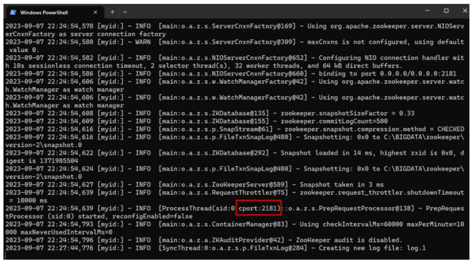
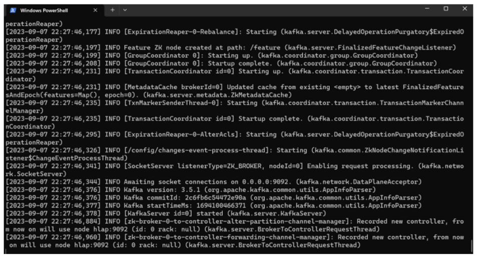


Figure 1. Deep Learning pipeline.



```
2023-09-07 22:24:54,578 [myid:] - INFO  [main:o.a.z.s.ServerCnxnFactory@169] - Using org.apache.zookeeper.server.NIOServ
erCnxnFactory as server connection factory
2023-09-07 22:24:54,580 [myid:] - WARN  [main:o.a.z.s.ServerCnxnFactory@309] - maxCnxns is not configured, using default
 value 0.
2023-09-07 22:24:54,582 [myid:] - INFO  [main:o.a.z.s.NIOServerCnxnFactory@652] - Configuring NIO connection handler wit
h 10s sessionless connection timeout, 2 selector thread(s), 32 worker threads, and 64 kB direct buffers.
2023-09-07 22:24:54,586 [myid:] - INFO  [main:o.a.z.s.NIOServerCnxnFactory@660] - binding to port 0.0.0.0/0.0.0.0:2181
2023-09-07 22:24:54,606 [myid:] - INFO  [main:o.a.z.s.w.WatchManagerFactory@42] - Using org.apache.zookeeper.server.watc
h.WatchManager as watch manager
2023-09-07 22:24:54,606 [myid:] - INFO  [main:o.a.z.s.w.WatchManagerFactory@42] - Using org.apache.zookeeper.server.watc
h.WatchManager as watch manager
2023-09-07 22:24:54,608 [myid:] - INFO  [main:o.a.z.s.ZKDatabase@135] - zookeeper.snapshotSizeFactor = 0.33
2023-09-07 22:24:54,609 [myid:] - INFO  [main:o.a.z.s.ZKDatabase@155] - zookeeper.commitLogCount=500
2023-09-07 22:24:54,616 [myid:] - INFO  [main:o.a.z.s.p.SnapStream@61] - zookeeper.snapshot.compression.method = CHECKED
2023-09-07 22:24:54,616 [myid:] - INFO  [main:o.a.z.s.p.FileTxnSnapLog@480] - Snapshotting: 0x0 to C:\BIGDATA\zookeeper\
version-2\snapshot.0
2023-09-07 22:24:54,622 [myid:] - INFO  [main:o.a.z.s.ZKDatabase@292] - Snapshot loaded in 14 ms, highest zxid is 0x0, d
igest is 1371985504
2023-09-07 22:24:54,624 [myid:] - INFO  [main:o.a.z.s.p.FileTxnSnapLog@480] - Snapshotting: 0x0 to C:\BIGDATA\zookeeper\
version-2\snapshot.0
2023-09-07 22:24:54,627 [myid:] - INFO  [main:o.a.z.s.ZooKeeperServer@589] - Snapshot taken in 3 ms
2023-09-07 22:24:54,639 [myid:] - INFO  [main:o.a.z.s.RequestThrottler@75] - zookeeper.request_throttler.shutdownTimeout
 = 10000 ms
2023-09-07 22:24:54,639 [myid:] - INFO  [ProcessThread(sid:0 cport:2181)::o.a.z.s.PrepRequestProcessor@138] - PrepReques
tProcessor (sid:0) started, reconfigEnabled=false
2023-09-07 22:24:54,793 [myid:] - INFO  [main:o.a.z.s.ContainerManager@83] - Using checkIntervalMs=60000 maxPerMinute=10
000 maxNeverUsedIntervalMs=0
2023-09-07 22:24:54,796 [myid:] - INFO  [main:o.a.z.a.ZKAuditProvider@42] - ZooKeeper audit is disabled.
2023-09-07 22:27:44,776 [myid:] - INFO  [SyncThread:0:o.a.z.s.p.FileTxnLog@284] - Creating new log file: log.1
```

Figure 2. Launched zookeeper successfully.



```
perationReaper)
[2023-09-07 22:27:46,177] INFO [ExpirationReaper-0-Rebalance]: Starting (kafka.server.DelayedOperationPurgatory$ExpiredO
perationReaper)
[2023-09-07 22:27:46,197] INFO Feature ZK node created at path: /feature (kafka.server.FinalizedFeatureChangeListener)
[2023-09-07 22:27:46,199] INFO [GroupCoordinator 0]: Starting up. (kafka.coordinator.group.GroupCoordinator)
[2023-09-07 22:27:46,208] INFO [GroupCoordinator 0]: Startup complete. (kafka.coordinator.group.GroupCoordinator)
[2023-09-07 22:27:46,231] INFO [TransactionCoordinator id=0] Starting up. (kafka.coordinator.transaction.TransactionCoor
dinator)
[2023-09-07 22:27:46,231] INFO [MetadataCache brokerId=0] Updated cache from existing <empty> to latest FinalizedFeature
sAndEpoch(features=Map(), epoch=0). (kafka.server.metadata.ZkMetadataCache)
[2023-09-07 22:27:46,235] INFO [TxnMarkerSenderThread-0]: Starting (kafka.coordinator.transaction.TransactionMarkerChann
elManager)
[2023-09-07 22:27:46,235] INFO [TransactionCoordinator id=0] Startup complete. (kafka.coordinator.transaction.Transactio
nCoordinator)
[2023-09-07 22:27:46,295] INFO [ExpirationReaper-0-AlterAcls]: Starting (kafka.server.DelayedOperationPurgatory$ExpiredO
perationReaper)
[2023-09-07 22:27:46,326] INFO [/config/changes-event-process-thread]: Starting (kafka.common.ZkNodeChangeNotificationLi
stener$ChangeEventProcessThread)
[2023-09-07 22:27:46,341] INFO [SocketServer listenerType=ZK_BROKER, nodeId=0] Enabling request processing. (kafka.netwo
rk.SocketServer)
[2023-09-07 22:27:46,344] INFO Awaiting socket connections on 0.0.0.0:9092. (kafka.network.DataPlaneAcceptor)
[2023-09-07 22:27:46,376] INFO Kafka version: 3.5.1 (org.apache.kafka.common.utils.AppInfoParser)
[2023-09-07 22:27:46,376] INFO Kafka commitId: 2c6fb6c54472e90a (org.apache.kafka.common.utils.AppInfoParser)
[2023-09-07 22:27:46,377] INFO Kafka startTimeMs: 1694100466371 (org.apache.kafka.common.utils.AppInfoParser)
[2023-09-07 22:27:46,378] INFO [KafkaServer id=0] started (kafka.server.KafkaServer)
[2023-09-07 22:27:46,884] INFO [zk-broker-0-to-controller-alter-partition-channel-manager]: Recorded new controller, fro
m now on will use node hlap:9092 (id: 0 rack: null) (kafka.server.BrokerToControllerRequestThread)
[2023-09-07 22:27:46,960] INFO [zk-broker-0-to-controller-forwarding-channel-manager]: Recorded new controller, from now
 on will use node hlap:9092 (id: 0 rack: null) (kafka.server.BrokerToControllerRequestThread)
```

Figure 3. Launched kafka successfully.

C. *Overview of NLP*

Natural Language Processing (NLP) – natural language processing, is a research branch of artificial intelligence (AI), focusing on the interaction between computers and natural human language, in the form of sound or text. NLP is divided into two major branches, including speech processing and text processing. These two branches are also closely related and complementary, such as converting text into speech and vice versa. Text processing is divided into two smaller branches, text understanding and text generation. NLP applications in practice are many, such as: automatic speech recognition (ASR), speech synthesis (text to speech – TTS, used in automatic text reading), information retrieval. information retrieval (IR, in search engines), question answering systems (QA), automatic text summarization (ATS), chatbots, machine translation (MT), testing Check spelling errors automatically.

Table I
HYPERPARAMTERS OF MACHINE LEARNING MODELS IN TEXT-BOX CLASSIFICATION APPROACH

Models	**Hyperparameters**
Logistic Regression	C: 100, penalty: l2
Multinomial Naive Bayes	alpha: 1.0
Support Vector Machine	kernel: rbf, gamma: scale, C: 10

D. *Overview of Deep Learning*

Deep Learning (DL) is a sub-branch of Machine Learning (ML), the birth of DL has promoted great progress in the field of artificial intelligence (AI). While machine learning is busy with labeled and unlabeled learning problems, deep learning continues to move forward by exploiting neural networks, which are descriptions of human nerves, creating the Strong in various artificial intelligence (AI) solutions such as computer vision, natural language processing, intelligent video analytics and many others . The development of deep learning cannot lack the influence of convolutional neural network (CNN) and recurrent neural network (RNN). Both machine learning and deep learning improve models through data, but the number of layers to train deep learning models is very large, that's why it is called ”deep learning”. Basic steps when solving a deep learning problem:

E. *Kafka & Spark*

Apache Kafka is a real-time distributed data streaming platform, capable of ingesting and processing trillions of records per day without latency, despite extremely large data volumes. Main components of Kafka: Producer, Consumer, Consumer group, Broker, Cluster, Zookeeper, Topic, Partitions, Zookeeper. In this experiment, we installed zookeeper and kafka on the laptop. [3]

Initially, to analyze and process big data, people used Hadoop because of its parallel programming and MapReduce architecture, flexible scalability, fault tolerance, and low cost. But the need for speed is a limitation of Hadoop, because almost everything is calculated on the hard drive. Apache Spark was born, overcoming that drawback, increasing calculation speed by several dozen to hundreds of times, and can be processed on RAM.

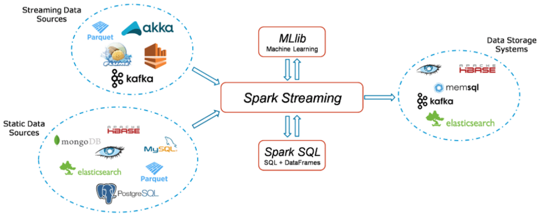


Figure 4. Spark architecture.

III. Dataset

We use selenium software to connect to the API of the tiki website, retrieve products in the selected category, and for each product, we will get the number of comment pages along with the rating. We conducted experiments on 410 products, each product received 10 pages of comments, receiving a total of 4,554 comments. And the second experiment was on 161 products, each product took 20 pages of comments and got 4,240 comments. The raw data is then processed by removing stop words, converting rating ≥ 4 into label 1, and rating ≤ 3 into label 0. Continue applying word embedding technique and split data into train and test sets, fed into models to evaluate accuracy.

IV. Evaluation

A. system design

Processing flow: ecommerce web, crawl data, put data into file or Kafka, load data (Spark/Pandas), BigDL (DL-lib)/Machine Learning/Deep Learning, Save in Mongo DB & display on Dashboard. The problem of classifying comments, basically according to the topic, is in the form of a classification problem. For the purpose of classifying comments as positive or negative, we can use multiclassification or binary classification, but for simplicity, in this topic we will use binary classification. The general model for binary classification is as follows. If Big Data techniques are applied to this problem, there will be a difference in the data collection and data transmission stages (Kafka), as well as the processing libraries applying Spark (BigDL) instead of Machine Learning/Deep Learning.

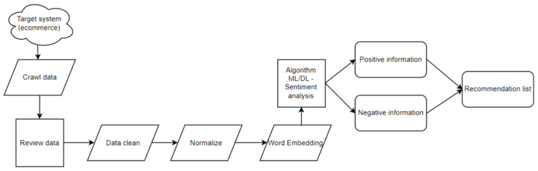


Figure 5. System architecture.

B. Experiment and evaluate the algorithm

Applying different algorithms to the two datasets, the results are compared through the table below. [4] [5]

The more comments on a product, the better results the training model will give. In addition, there is no difference between the algorithms through accuracy and confusion matrix. Comments based on 1 product have an imbalance, leading to an impact on the training model.

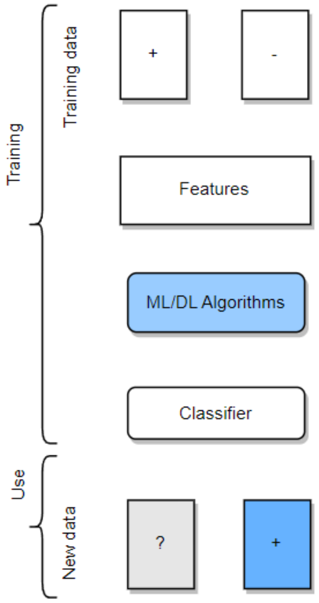


Figure 6. Comment classification application model.

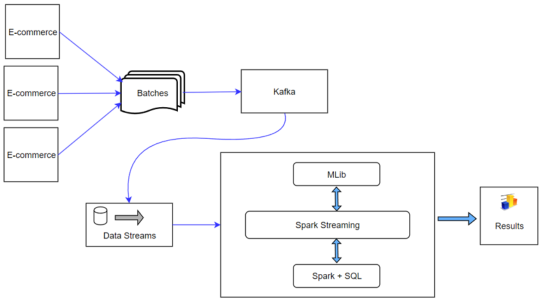
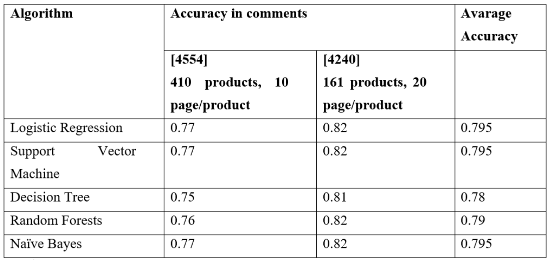


Figure 7. Big Data application system design model.



Algorithm	Accuracy in comments		Avarage Accuracy
	[4554] 410 products, 10 page/product	[4240] 161 products, 20 page/product	
Logistic Regression	0.77	0.82	0.795
Support Vector Machine	0.77	0.82	0.795
Decision Tree	0.75	0.81	0.78
Random Forests	0.76	0.82	0.79
Naïve Bayes	0.77	0.82	0.795

Figure 8. Accuracy according to comments.

V. Conclusion

A. Conclusion

With limited time and scale of the project, we just stopped at experimentally applying Machine Learning, Deep Learning, and Big Data algorithms to handle the problem of classifying comments on e-commerce

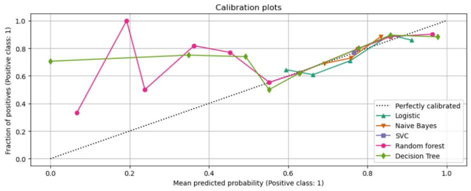


Figure 9. The distribution of the mean value of the prediction.

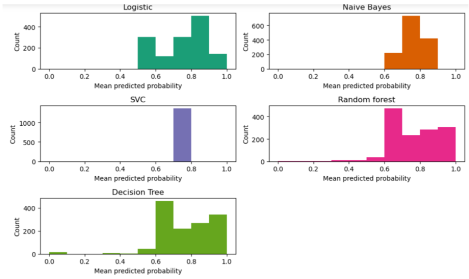


Figure 10. The results are distributed according to histogram based on algorithms.

websites. For simplicity, and also to suit the current business needs of businesses, try to reduce the actual problem to the form of a binary classification problem. Analyzing a system, be it an e-commerce site, social network or a similar system, the data collection stage depends entirely on that system, which may change over time, depending on the system. The technology that the system uses has different approaches. According to business operations, comment analysis is just a factor for reference, helping business planners or managers and business owners have more perspectives to support appropriate decision-making. Also note that this is only a means to provide reference information, there is still a need for a career in the analytical business to not lead to misleading policy decisions. This topic can be developed into an early warning system of brand risks for businesses, based on understanding trends on the internet and social networks in advance to prepare for businesses in advance. However, connecting multiple systems that need to be analyzed requires a professional and meticulous operations team to be ready to fine-tune and regularly check the analyzed data for general trends, which is not easy. easy. In addition, the system needs to be accessed quickly, data may have to be stored for several periods to be analyzed at the same time, requiring a certain level of hardware investment.

B. Future works

In the next steps, Although the problem of comment analysis has been studied by many groups of authors before, this thesis synthesizes different methods to handle it, trying to improve accuracy, visualize data and apply techniques. Big Data techniques aim to create a new aspect in research. The method implemented in this thesis can be fully applied to develop (1) a brand risk prediction system, (2) a trend analysis system based on multi-faceted user comments, (3) market trend collection and analysis system applicable to market research companies or marketing departments. In addition, the data collection method in the thesis can also be applied to collect data for many other systems, especially the stream data transmission method will be useful for string problems or data processing.

References

[1] D. Phuc, “Social network and application analysis.” Ho Chi Minh, VN: Ho Chi Minh City National University Publishing House, 2017.

[2] “Hate speech detection on vietnamese social media text using the bi-gru-lstm-cnn model.” University of Information Technology of Ho Chi Minh City, VietNam, 2019.

[3] D. T. Hop, “Big data, smart software system laboratory.” Ho Chi Minh, VN: Ho Chi Minh City National University Publishing House, 2020.

[4] I. Dr. Mary Vennila S, Raviya K, “An implementation of hybrid enhanced sentiment analysis system using spark ml pipeline: A big data analytics framework.” International Journal of Advanced Computer Science and Applications, Oct.-Nov. 2021.

[5] M. Molaei and D. Mohammadpur, “Distributed online preprocessing framework for big data sentiment analytics.” Department of Computer Engineering, University of Zanjan, Iran, Journal of Artificial Intelligence and Data Mining (JAIDM), Oct.-Nov. 2022, pp. 197–205.