[1] **Irvan Santoso1**

[2] **Edi Abdurachman**

[3] **Harco Leslie Hendric Spits Warnars**

[4] **Lili Ayu Wulandhari**

Development of a Model to Detect the Validity of Indonesian Reviews on E-Commerce Products Using Bert and Smart Approaches

Abstract: **-** Ensuring customer satisfaction in e-commerce relies heavily on the accuracy of product reviews. The validity of reviews is critical as accurate and reliable reviews significantly influence consumer decision-making. An effective method to ensure the validity of reviews is through sentiment analysis, specifically on open-ended questions, and comparing them with Likert Scale values. This approach helps in identifying inconsistent or manipulative reviews and provides deeper insights into overall customer satisfaction. This study used BERT and SMART approaches to achieve high accuracy in consumer feedback sentiment analysis. The results showed that IndoBERT, among the various methods, produced the highest accuracy compared to BERT, DistilBERT, ALBERT, and RoBERTa. Notably, combining IndoBERT with SMART achieved the best overall accuracy, outperforming other combinations. Although SMART slightly improved accuracy by around 1%, further research is needed to evaluate the impact on processing time of this approach.

Keywords— BERT, SMART, Sentiment Analysis, IndoBERT, E-Commerce.

I. INTRODUCTION

Undoubtedly, information technology plays a pivotal role in advancing the field of business. A tangible example in the business world is the establishment of e-commerce [1]. According to researched data, one of the top e-commerce platforms in Indonesia had an average of 107.4 million visitors during Q1-Q3 of 2023 [2]. With such a large number of visitors, e-commerce faces challenges in maintaining customer engagement [3]. One critical aspect to sustain customer engagement is ensuring customer satisfaction with the e-commerce system and the information provided [4].

However, maintaining satisfaction with the available information poses its own difficulties and presents new challenges that must be addressed [5]. One such challenge relates to product reviews on e-commerce platforms [6]. Consumers sometimes hesitate to purchase desired products due to uncertainties about the validity of reviews and ratings given by other consumers [7]. Moreover, accurately reflecting customer sentiment towards products remains a challenge, crucial for sellers in improving service quality and customer satisfaction [8]. Manual collection of free-text review data with Likert scale scoring can lead to inaccuracies in assessing quality, due to differences between Likert scores and the actual content of user reviews, whether intentional or not by users [9].

Therefore, this study aims to develop an automated sentiment analysis model to help validate customer reviews on e-commerce platforms. The model will categorize sentiments in text, such as product, service, and delivery into positive, negative, or neutral sentiments, then compare them with the Likert scale ratings provided by users to determine review validity. Additionally, the study will utilize approaches like Bidirectional Encoder Representations from Transformers (BERT) [10] and Smoothness-inducing Adversarial Regularization and BRegman pRoximal poinT opTimization (SMART) [11], previously developed by researchers.

Following the classification of reviews, validation will proceed by matching them against customer Likert ratings. Furthermore, the automatically obtained results will be compared with manual judgments to ensure the accuracy of the developed model. Validating reviews in this manner will serve as a tool to enhance customer trust, thereby contributing to increased satisfaction with managed e-commerce platforms.

[1] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480. isantoso@binus.edu
[2] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480. ediabdurachman@gmail.com
[3] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480. Spits.hendric@binus.ac.id
[4] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480. lili.wulandhari@binus.ac.id

II. PREVIOUS STUDY

In 2018, Devlin et al. [12] conducted a study introducing Bidirectional Encoder Representations from Transformer (BERT), a novel representation model. BERT performs deep bidirectional training on unlabeled text, leveraging contextual clues across layers. The research underscores the significance of bidirectional pre-training for language representation and illustrates how pre-trained models reduce the reliance on complex architectures.

Liu et al. [13] conducted a study titled "RoBERTa: A Robustly Optimized BERT Pretraining Approach," which aimed to measure the impact of various parameters and different forms and amounts of data. The data used comprised BookCorpus, with 800 million entries, and English Wikipedia, with 2.5 billion entries, totaling 16 GB. The key contribution of this research was the development of a process to determine alternative BERT training designs, which improved the performance of downstream tasks compared to other models. Additionally, Liu et al. highlighted that using larger datasets during pre-training can enhance performance on downstream tasks.

Similarly, Lan et al. [14] conducted research titled "ALBERT: A Lite BERT For Self-Supervised Learning of Language Representations," which aimed to improve training speed using BERT. In their study, A Lite BERT (ALBERT) was designed with an architecture featuring fewer parameters compared to BERT. This reduction allowed ALBERT to significantly shorten pre-training time. The study's contributions include optimizations in factorized embedding parameterization, cross-layer parameter sharing, and inter-sentence coherence loss. Moreover, the data used in this research also included BookCorpus and English Wikipedia, consistent with the data used in BERT studies.

Furthermore, Wilie et al. [15] extended BERT for Indonesian language by adapting it to structural requirements. Indonesian's distinct linguistic structure influences natural language processing outcomes across diverse languages. Data for their study were gathered from blogs, news, and websites, showing that IndoNLU achieves superior accuracy in Indonesian language processing, guiding future research. Additionally, other researchers explored methods to enhance sentiment analysis results. Jiang et al. [16] tackled overfitting during fine-tuning by proposing the SMART model, integrating smoothness-inducing regularization [17] and Bregman proximal point optimization [18].

III. DATA COLLECTION

The dataset obtained is consumer reviews of products and compiled manually by combining datasets from Kaggle, GitHub, and directly from one of the largest marketplaces in Indonesia. The reviews used in the dataset are entirely in Indonesian. The total dataset consists of 12,722 entries, consisting of free text reviews and Likert scores.

After the dataset is obtained, preprocessing is carried out to prepare the raw data for easy use in the next process. Furthermore, the data will be divided into training, validation, and testing sets. This division will be carried out with a ratio of 70% for training, 15% for model validation, and 15% for model testing. This dataset division is important to reduce the risk of overfitting and underfitting.

IV. METHODOLOGY

There are several processes for detecting the validity of customer reviews, which will be explained in Figure 1. Figure 1 provides an overview of the process of developing a review validity detection model, referring to the sentiment analysis process starting from data cleaning and category labeling before entering pre-processing.

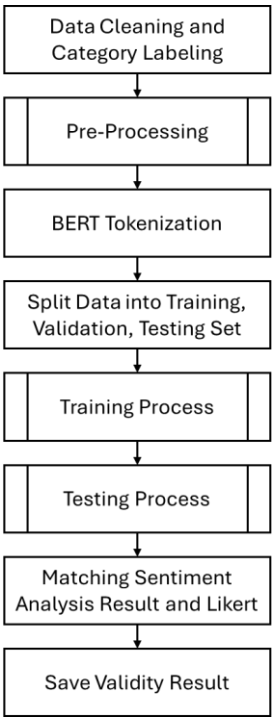

Fig. 1. Model of Validity Check

Data cleaning is essential to enhance the quality and consistency of the dataset used, aiming to avoid noise in the data, such as removing irrelevant elements like unclear characters such as "bagusjnsneiuenie" and eliminating emoticons. Table 1 illustrates a comparison between raw review data and cleaned reviews.

Table 1: Data Cleaning Example

Reviews	**Data Cleaning**
barangnya bagus bgtrnnefunesiunsunskujni…….	barangnya bagus bgt……
Gila sih bagus banget barangnya, gambar sama nyatanya bagusnya sama 😊. ga cuma buat keyboard mechanical ini, buat Midi Controller Korg NanoKontrol 2 juga masuk keren banget 😎😎 jadi secara ngak langsung bisa jadi gig bag case buat midi controller 👍👍👍👍👍	Gila sih bagus banget barangnya, gambar sama nyatanya bagusnya sama . ga cuma buat keyboard mechanical ini, buat Midi Controller Korg NanoKontrol 2 juga masuk keren banget jadi secara ngak langsung bisa jadi gig bag case buat midi controller

Furthermore, label categorization and assessment of each existing review will be carried out. The categories themselves are divided into three categories, namely product, service, and delivery. Meanwhile, for the review score, a Positive (Pos), Negative (Neg), or Neutral (Neu) value will be given. Categorization and assessment of reviews will be carried out manually first as comparison material against the results issued by the developed model. Examples of categorization and giving sentiment values will be illustrated in Table 2 and Table 3.

Table 2: Category Explanation

Category	**Description**
Product (P)	Reviews containing sentiments about related products
Service (S)	Reviews containing sentiments regarding the service provided by the seller
Delivery (D)	Reviews containing sentiments related to the delivery service process

Table 3: Example of Sentiment per Category

ID	**EN**	**P**	**S**	**D**
Produknya bagus, kualitas baik, tapi penjual kurang responsive	The product is good, the quality is good, but the seller is not very responsive	Pos	Neg	Neu

After undergoing the preprocessing stage, the data will proceed to the BERT tokenization process, where sentences are converted into tokens understandable by the IndoBERT model. Following BERT tokenization, the data will be divided into three sets: training, validation, and testing sets, which will be used during model training, evaluation, and testing phases.

The next stage is the training process, involving training the model using the training data to learn patterns and characteristics of the review data. This is followed by the testing process, where the trained model is evaluated using the testing data to assess its performance and generate sentiment analysis results.

Subsequently, the sentiment analysis results obtained from the training process will be used in the Load Sentiment Analysis Result stage to obtain sentiment analysis results. These results will then be compared with the Likert scores in the Matching Sentiment Analysis Result and Likert stage to determine sentiment validity.

The final step is Save Validity Result to Excel, where the validity results will be stored in Excel format for documentation and further analysis purposes. Figure 2 provides an overview of the preprocessing process, which includes several stages.

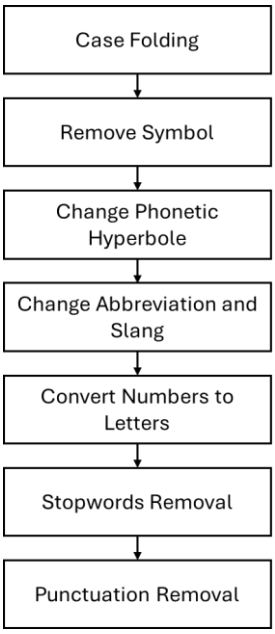

Fig. 2. Pre-processing Process

The preprocessing process begins with Case Folding, which converts all uppercase letters in sentences to lowercase. Next is the Remove Symbol process to eliminate all unnecessary symbols and punctuation marks from the text. Following that, the Change Phonetic Hyperbole process removes letters with phonetic exaggerations, such as excessive letter repetitions, and standardizes them. This is followed by Change Abbreviation & Slang to convert abbreviations and slang into their full or standard forms for better understanding. The next step is Convert Numbers to Letters, where numbers in the text are converted into letters. Then, the Stopwords Removal process removes common words that do not significantly contribute to the analysis, such as "dan" (and), "di" (in), "atau" (or), and "yang" (that/which). Afterward, punctuation removal is performed to eliminate punctuation marks that do not hold value in sentiment analysis processes.

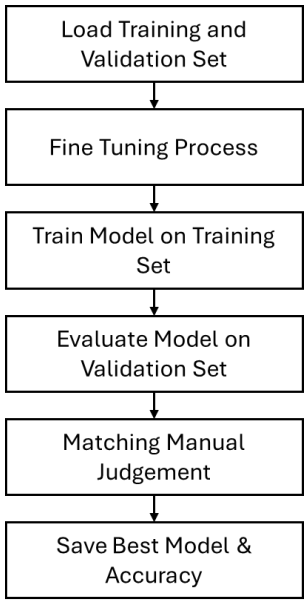

Fig. 3. Training Process

Based on Figure 3, the model training process begins with loading the necessary training and validation datasets to train and evaluate the model. Once the training data is loaded for processing, the fine-tuning stage of model parameters will be manually conducted by the author. Following this, the model training process uses the training dataset to train the model, evaluates its performance using the validation dataset, and compares the model's predictions with Manual Judgments made by the author in previous stages to obtain accuracy scores.

Subsequently, the model's hyperparameter values and accuracy scores will be stored in JSON format for comparison purposes. The best accuracy score will be saved as the best model, which will then be used for validation and testing processes. Figure 4 depicts the flowchart related to the model testing process.

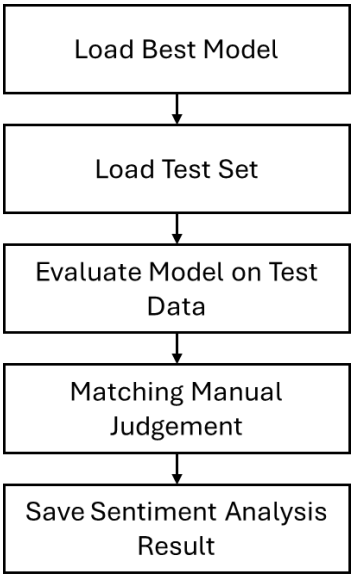

Fig. 4. Testing Process

Figure 4 depicts the process of testing a trained sentiment analysis model. This testing process begins by utilizing the previously trained best model. Next, unseen test data, which the model has not encountered during training, is used to evaluate the model. The model evaluation process calculates the accuracy of the model by comparing its predictions with manual assessments. After computing accuracy and other performance metrics, the sentiment analysis results are saved for the validity check process.

V. RESULT AND DISCUSSION

Based on the experiments conducted, the results and discussions for several key processes will be presented. Table 4 below shows an example of the data pre-processing results.

Table 4: Example of Pre-Processing Process

Reviews	Pre-Processing Result
Pesen di tanggal 20, hari itu juga langsung diproses sama penjual ke ekspedisi dan hari itu juga dikirim sama ekspedisi, mantep banget deh memang', langsung sampe di tanggal 22	pesan tanggal dua puluh hari langsung proses sama jual ekspedisi hari kirim sama ekspedisi mantap banget memang langsung tanggal dua puluh dua
pengiriman lama…...:"""">:"<">">, respon penjual tidak ramah	kirim lama respon jual tidak ramah
Tidak ada upaya dari penjual untuk menyelesaikan masalah saya dengan cepat meskipun udh berkali kali dihubungi, tapi MASIH GAK ADA tindakan nyata!! Proses pengirimannya juga kurang baik sangat lemot.	tidak jual selesai cepat meski sudah kali hubung tidak tindak nyata proses kirim kurang baik sangat lambat

Based on the results obtained from the pre-processing process, the sentiment analysis process for open questions will be continued using the BERT approach. The training of the BERT model will involve integrating the SMART approach after undergoing preprocessing stages. This training process focuses on optimizing model parameters to achieve optimal performance and predictions, known as the best model. The approaches that will be utilized include BERT, DistilBERT, ALBERT, RoBERTa, and IndoBERT, which will also be tested with the SMART approach.

Table 5: Training Result

Approach	Accuracy per Category		
	Product	**Service**	**Delivery**
BERT	0.9589	0.9589	0.9558
DistilBERT	0.9498	0.9491	0.9452
ALBERT	0.9480	0.9481	0.9459
RoBERTa	0.9432	0.9433	0.9402
IndoBERT	0.9631	0.9629	0.9580
BERT–SMART	0.9670	0.9671	0.9632
DistilBERT–SMART	0.9601	0.9590	0.9543
ALBERT–SMART	0.9577	0.9569	0.9537
RoBERTa–SMART	0.9522	0.9540	0.9514
IndoBERT–SMART	0.9740	0.9732	0.9701

As shown in Table 5, the training results for all approaches exhibit good accuracy, with values above 94%. However, among all approaches without using SMART, IndoBERT achieved the highest accuracy with 96.31% for the Product category, 96.29% for the Service category, and 95.8% for the Delivery category. When combining all BERT approaches with SMART, IndoBERT-SMART achieved the highest accuracy with 97.4% for the Product category, 97.32% for the Service category, and 97.01% for the Delivery category. This indicates that IndoBERT is the most effective approach for processing the data. Additionally, IndoBERT is specifically designed for processing Indonesian sentiment, which contributes to its superior performance with Indonesian language data.

Next, the sentiment values obtained will be compared with those from manual judgment. The accuracy of these comparisons will be presented in Table 6.

Table 6: Matching Training Result to Manual Judgement Result

Approach	Accuracy per Category		
	Product	**Service**	**Delivery**
BERT	0.8977	0.8972	0.8904

Approach	Accuracy per Category		
	Product	**Service**	**Delivery**
DistilBERT	0.8871	0.8869	0.8830
ALBERT	0.8868	0.8867	0.8843
RoBERTa	0.8842	0.8811	0.8822
IndoBERT	0.9012	0.9010	0.8999
BERT–SMART	0.9055	0.9049	0.9015
DistilBERT–SMART	0.8970	0.8965	0.8940
ALBERT–SMART	0.8948	0.8947	0.8922
RoBERTa–SMART	0.8921	0.8901	0.8909
IndoBERT–SMART	0.9105	0.9100	0.9095

Based on Table 6, there is a decline in accuracy when performing manual judgment matching. However, the results still achieve good values, with IndoBERT maintaining the highest accuracy without using SMART, recording 90.12% for the Product category, 90.1% for the Service category, and 89.99% for the Delivery category. With the SMART approach, IndoBERT achieved 91.05% accuracy for the Product category, 91% for the Service category, and 90.95% for the Delivery category.

Following the training results, the next step will involve testing all approaches and subsequently performing manual judgment matching. The results of this testing will be detailed in Tables 7 and 8.

Table 7: Testing Result

Approach	Accuracy per Category		
	Product	**Service**	**Delivery**
BERT	0.9577	0.9575	0.9571
DistilBERT	0.9489	0.9488	0.9479
ALBERT	0.9482	0.9482	0.9478
RoBERTa	0.9433	0.9437	0.9425
IndoBERT	0.9630	0.9631	0.9629
BERT–SMART	0.9679	0.9670	0.9673
DistilBERT–SMART	0.9572	0.9570	0.9570
ALBERT–SMART	0.9565	0.9567	0.9561
RoBERTa–SMART	0.9530	0.9535	0.9529
IndoBERT–SMART	0.9722	0.9721	0.9722

Table 8: Matching Testing Result to Manual Judgement Result

Approach	Accuracy per Category		
	Product	**Service**	**Delivery**
BERT	0.8995	0.8992	0.8993
DistilBERT	0.8854	0.8855	0.8843
ALBERT	0.8849	0.8844	0.8841
RoBERTa	0.8802	0.8805	0.8799
IndoBERT	0.9026	0.9026	0.9024
BERT–SMART	0.9046	0.9043	0.9041
DistilBERT–SMART	0.8968	0.8964	0.8965
ALBERT–SMART	0.8946	0.8944	0.8944
RoBERTa–SMART	0.8919	0.8922	0.8918
IndoBERT–SMART	0.9103	0.9102	0.9102

Based on Tables 7 and 8, the results obtained are similar to those from the training and manual judgment matching processes. The testing results indicate that IndoBERT continues to outperform other similar approaches. IndoBERT, without integrating the SMART approach, achieved accuracy rates of 96.3% for the Product category, 96.31% for the Service category, and 96.29% for the Delivery category. When combined with SMART, IndoBERT achieved even higher accuracy rates of 97.22% for the Product category, 97.21% for the Service category, and 97.22% for the Delivery category.

In comparison to manual judgment matching, IndoBERT without SMART achieved accuracy rates of 90.26% for the Product category, 90.26% for the Service category, and 90.24% for the Delivery category. With the SMART approach, IndoBERT attained accuracy rates of 91.03% for the Product category, 91.02% for the Service category, and 91.02% for the Delivery category.

VI. CONCLUSIONS

Maintaining customer satisfaction in e-commerce heavily relies on the accuracy of product reviews. The validity of reviews is crucial as accurate and trustworthy reviews significantly influence consumer decision-making. An effective method to ensure review validity is through sentiment analysis, particularly on open-ended questions, and comparing it with Likert Scale values. This approach not only helps in identifying inconsistent or manipulative reviews but also provides deeper insights into overall customer satisfaction.

The study demonstrated that using BERT and SMART approaches yielded high accuracy in determining the sentiment of consumer-provided sentences. The results from manual judgment matching also showed good accuracy and can be used as a reference for assessing the validity of product reviews. Among various approaches, IndoBERT emerged as the best for the specified dataset, outperforming BERT, DistilBERT, ALBERT, and RoBERTa in accuracy. Notably, BERT still delivered better results on the prepared dataset compared to DistilBERT, ALBERT, and RoBERTa, despite their subsequent developments.

Furthermore, when combined with the SMART approach, IndoBERT-SMART achieved the highest accuracy compared to BERT-SMART, DistilBERT-SMART, ALBERT-SMART, and RoBERTa-SMART. BERT-SMART also outperformed other approaches in accuracy. This indicates that the SMART approach can enhance the accuracy of BERT, although the improvement is relatively modest, around 1% or less. Future research should consider evaluating the time efficiency of using SMART, as it may result in longer processing times.

REFERENCES

[1] Laudon, K. C., & Traver, C. G. (2020). E-commerce 2019: Business, technology, society. Pearson.

[2] Ahdiat, A. (2023). Pengunjung Shopee dan Blibli Naik pada Kuartal II 2023, E-Commerce Lain Turun. (Online), diakses 15 Januari 2024 dari https://databoks.katadata.co.id/datapublish/2023/07/07/pengunjung-shopee-dan-blibli-naik-pada-kuartal-ii-2023-e-commerce-lain-turun.

[3] Gupta, A. (2014). E-Commerce: Role of E-Commerce in today's business. International Journal of Computing and Corporate Research, 4(1), 1-8.

[4] Gajewska, T., Zimon, D., Kaczor, G., & Madzík, P. (2020). The impact of the level of customer satisfaction on the quality of e-commerce services. International Journal of Productivity and Performance Management, 69(4), 666-684.

[5] Ingaldi, M., & Ulewicz, R. (2019). How to make e-commerce more successful by use of Kano's model to assess customer satisfaction in terms of sustainable development. Sustainability, 11(18), 4830.

[6] Nisar, T. M., & Prabhakar, G. (2017). What factors determine e-satisfaction and consumer spending in e-commerce retailing?. Journal of retailing and consumer services, 39, 135-144.

[7] Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE access, 8, 23522-23530.

[8] Sharma, G., & Lijuan, W. (2015). The effects of online service quality of e-commerce Websites on user satisfaction. The electronic library, 33(3), 468-485.

[9] Santoso, I., Abdurachman, E., Warnars, H. L. H. S., & Wulandhari, L. A. (2023). MODEL DEVELOPMENT OF AUTOMATIC VALIDATION DETECTION ON SURVEY RESPONSES OF COURSE LEARNING PROCESSES USING BERT-BASED MODEL. Journal of Research Administration, 5(2), 12871-12884.

[10] Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

[11] He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[13] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[14] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[15] Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., ... & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. arXiv preprint arXiv:2009.05387.

[16] Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Zhao, T. (2019). Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. arXiv preprint arXiv:1911.03437.

[17] Miyato, T., Maeda, S. I., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1979-1993.

[18] Eckstein, J. (1993). Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. Mathematics of Operations Research, 18(1), 202-226.