






Application of Term Frequency - Inverse Document Frequency in The Naive Bayes Algorithm For ChatGPT User Sentiment Analysis

Novita Rajagukguk¹^{*}, I Putu Eka Nila Kencana²
I GN Lanang Wijaya Kusuma³

^{1,2,3}Udayana University, Bali, Indonesia
*novitarajagukguk920@gmail.com

Abstract.

Sentiment analysis as part of Natural Language Processing has been widely used to see public sentiment towards a topic. Sentiment analysis functions to classify opinions into positive or negative classifications. In classifying opinions, an algorithm is needed to manage opinion data. One well-known algorithm capable of classifying text data simply and accurately is the naïve Bayes algorithm. Therefore, this research will use the Naive Bayes algorithm which can work well on high-dimensional data. The valid data used in this research is 36,000 ChatGPT user reviews from the Google Play Store, while the outsample data used is 400 tweets from X application users. To increase the classification accuracy value, the naïve Bayes algorithm is accompanied by feature weighting using the Term Frequency-Inverse Document Frequency technique. The performance of the classification model shows an accuracy value of 84%, recall of 84%, and precision of 83%. Next, the model classification is stored in pickled form and used to predict outsample data. The predicted data shows data with 208 negative labels and 192 positive labels.

Keywords: Sentiment Analysis, Naïve Bayes, Term-Frequency.

1 Introduction

Natural Language Processing (NLP) is widely used to solve various human problems. NLP is stated as a theoretically motivated set of computer techniques used to analyze and represent languages, genres, and modes at various levels of linguistic analysis with the ultimate goal of being able to complete a variety of human tasks [1]. One example of applying NLP is the ChatGPT application which can help humans answer questions via a chat system. ChatGPT has become a hot topic of conversation because of its usefulness. Therefore, knowing public opinion regarding ChatGPT can be used to evaluate and develop ChatGPT applications.

In studying people's opinions on a topic, there is the term sentiment analysis. Sentiment analysis or opinion mining is a field of science that studies and analyzes opinions, sentiments, attitudes, and emotions regarding a topic. In sentiment analysis, the program will classify an opinion into positive or negative classifications [2]. In classifying

opinions, an algorithm is needed to manage opinion data. One well-known algorithm that can classify text data simply and accurately is the Naïve Bayes algorithm.

The Naïve Bayes (NB) algorithm is a classification algorithm with a probabilistic classification technique which is famous because it is simple and accurate[3]. In Rish research (2001) said that the accuracy of NB does not depend directly on the level of feature dependence [4]. However, the research of Renie et al. (2003) said that using NB for independent features will result in low accuracy. The application of NB in classifying text data which shows that the features in the data tend to be related to each other tends to provide low accuracy. In conditions like this, NB will choose bad weights in determining decision boundaries [5].

To overcome the problem of interdependent features, appropriate weight adjustments are required [5]. One of the weighting features that can be used is the Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a method that assigns weights to tokens based on the frequency of tokens appearing in the document. TF-IDF is used to view data with a high level of relevance to the label [6].

A lot of research has been done on sentiment analysis. One of the studies regarding sentiment analysis was conducted by Jain (2023) who compared several machine learning and deep learning algorithms to analyze X user sentiment regarding Covid-19 vaccination. The research results show that the model performance metric with the highest accuracy uses machine learning with the Support Vector Machine (SVM) algorithm, namely 88.7989%, while in deep learning the highest accuracy uses the Long Short Term Memory algorithm with an accuracy of 90.42% [7].

Another study by Bahassine et al (2018) used Chi-Square and Information Gain feature selection to increase the accuracy of Arabic text classification results. By using the SVM algorithm, higher accuracy results were obtained in Chi-Square feature selection, namely 90.50% [8].

In 2020, Atmadja, et al conducted a sentiment analysis of opinions regarding online transportation on social media. This research used 565 tweets data which was divided into 500 as training data and 65 as test data. The research results show that the accuracy value using the NB algorithm is 66.15% and for the K-Nearest Neighbors algorithm the accuracy is 67.69% [9].

To find out public opinion regarding public acceptance of the new normal rules after COVID-19, Samsudin et al conducted sentiment analysis research using the NB algorithm. The data used in this research was 2807 tweet data with research evaluation showing that NB can classify data with 83% accuracy [10].

Then, Sarasvananda et al. conducted a sentiment analysis of Twitter users' opinions regarding online learning in Indonesia. This research was followed by weighting words using TF-IDF method. By using the NB algorithm, an accuracy value of 99.87% was obtained [11].

From the results of previous research, many researchers use the NB algorithm with feature weighting using the TF-IDF method with fairly high accuracy results. Therefore, research into ChatGPT user sentiment analysis using the NB algorithm is feasible.

2 Methods

2.1 Dataset

This research uses data obtained from a review of the ChatGPT application found in the Play Store application. The ChatGPT application review data will be referred to as valid data. Valid data was obtained by scrapping data using the Google Play Scrapper library in Python software. Then the data that will be used as outsample data is social media user X's tweet data regarding ChatGPT. Outsample data was collected using the Nitter library which is available in Python.

2.2 Pre-processing

Data pre-processing is the stage of cleaning data that is still filled with signs or symbols that cannot be understood by the computer so that the computer can understand it. The pre-processing stage consists of several stages, namely:

Case Folding

The case folding stage is carried out to change the letters in the dataset to lowercase with the aim that the text data to be processed has the same case. The use of lowercase and uppercase letters can affect the accuracy of the model because it will assume that tokens with the same meaning are two different things. In this research, the data will be converted to lowercase because it is more practical than having to convert it to uppercase [12].

Cleaning

Cleaning is the stage where text data will be cleaned from numbers, punctuation, special marks, URLs, and other special characters. The cleaning stage is useful for improving classification accuracy because special signs will be considered as noise [13].

Tokenization

Tokenization is the stage where a sentence is broken down into words, characters, or punctuation. In the tokenization stage, the results of sentence solving are called tokens [14].

Stopwords

The next step is stopwords to filter out words that are considered to not affect the sentiment analysis results. The stopwords stage will discard tokens that are considered to provide little information for the classification process [12].

Stemming

The stemming stage will change each word into a base word. Stemming is needed so that each form of the same word has the same case so that it does not give rise to double meanings [12].

2.3 Feature Extraction

Feature extraction is the process of giving word weights using tokens that have been pre-processed. One data extraction technique is the Term Frequency-Inverse Document Frequency (TF-IDF) which looks at how often words appear in documents [6]. TF-IDF will convert text into vector form so that it can be processed by a computer [15]. The TF-IDF method is a combination of Term Frequency (TF) and Inverse Document Frequency (IDF).

TF states the frequency with which a term appears in a document. The more often a term appears in a document, the greater its weight. In this case, $f(t_k, d_j)$ shows the number of occurrences of term k in a document j . The TF formula is:

$$TF(t_k, d_j) = f(t_k, d_j) \quad (1)$$

IDF shows that if a term appears frequently in many documents, the IDF value will be smaller. IDF itself can be used to distribute terms in a document collection. The formula for IDF is:

$$IDF(t_k) = \log \frac{D}{df(t)} \quad (2)$$

from the TF and IDF formulas, TF-IDF has the form, namely:

$$TF - IDF(t_k, d_j) = TF(t_k, d_j) \times IDF(t_k) \quad (3)$$

with (t_k, d_j) stating the number of occurrences of term k in a document j , D stating the number of documents in the dataset, and $df(t)$ stating the number of documents containing the term.

2.4 Feature Selection

In text classification, feature selection is important to select the most informative features and provide the highest contribution to the target class. This research will use a filter method with the Chi-Square formula to calculate features and select features to increase the accuracy of classification results. In text classification, Chi-Square will perform an independence test on features and then test whether there is a relationship between a term and a category during the feature selection process [16]. In mathematical form, the Chi-Square test can be written as follows:

$$\text{Chi - Square } (\chi^2) = \sum \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (4)$$

where A_{ij} is the observation data value and E_{ij} is the expected value of A_{ij} . If a feature is found that approaches more classes, that feature will give a higher feature score so that it can be generalized to all classes.

2.5 Naïve Bayes

Naïve Bayes (NB) is a Bayesian classification approach based on Bayes' theorem to predict the probability of a certain set of features as part of a label. NB algorithm uses conditional probability in its application, where event A will occur if the individual probabilities of A and B and the conditional probability of B occurring are known [17]. Mathematically, conditional probability can be written as follows:

$$P(Y = y|X = x) = \frac{\prod_{i=1}^n P(X=x_i | Y=y) P(Y=y)}{\sum_i^C P(y_i, X=x)} \quad (5)$$

Let X_1, X_2, \dots, X_n be a feature of the D_x domain, where $\{x_1, x_2, x_3, \dots, x_n\}$ is the observed value. Where Y is the target feature from the domain $D_y = \{0,1\}$. It is assumed that there is a D_x to D_y function. In this case, NB aims to choose the class Y that maximizes the posterior probability, $P(Y = y | X = x)$. $P(Y = y)$ is the prior probability and $P(X = x | Y = y)$ is the conditional probability of the target feature (class). Assuming that the features are conditionally independent then $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | Y = y) = \prod_{i=1}^n P(X = x_i | Y = y)$. In practice, NB doesn't care about $P(X = x)$. because its value will be the same for each class.

In the real world, the data found may not follow the NB assumption, namely independent features. If the data has features that are dependent and still use NB, then classification will show poor results and hidden patterns will be difficult to extract. Therefore, it is necessary to use a NB approach to overcome the problem of features that may depend. An approach that can be taken is to use reduced conditional independence assumptions known as Semi-Naïve Bayes and use weighting features to increase the influence of features [18].

2.6 Evaluation

The evaluation stage can be carried out using cross validation to obtain accurate and stable estimates. Furthermore, this research will use k-fold cross validation to divide the dataset into training subsets and testing which is carried out repeatedly according to the specified k value. Model performance will be measured with accuracy, recall and precision values can be calculated [19].

Accuracy measures the predicted value according to reality compared to the overall data. In mathematics, accuracy can be written in equation (6).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Recall to measure a lot of data that is predicted to be positive compared to all data that is essentially positive. In mathematics, recall can be written in equation (7).

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

Precision measures the amount of data that is predicted to be positive compared to data that is predicted to be positive without considering the truth value of the prediction in reality. In mathematics, precision can be written in equation (8).

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

3 Road Map

3.1 Collecting Data

This research uses data obtained from reviews of the ChatGPT application found in the Play Store website with a dataset of 36,000 ChatGPT user opinions. Review data taken from the Google Play Store website is the most relevant data when data is collected. In this research, the data collected consists of rating and review attribute user reviews regarding ChatGPT. The data reviews consist of two classes, namely the positive class and the critical thinking class which will be used to build model classification. The data that will be used as outsample data is the review data of social media user X regarding ChatGPT as many as 400 opinions of user X regarding ChatGPT.

3.2 Preprocessing Data

The pre-processing stage will be implemented for valid data and sample data. The data will go through the case folding, cleaning, tokenization, stopwords, and stemming stages

3.3 Feature Extraction

The dataset that has been cleaned in the pre-processing stage will then be given word weights using the TF-IDF by using the libraries available in Python, the output of the feature extraction stage is a token that has been transformed into a vector.

3.4 Feature Selection

The input used in performing feature selection is the output of feature extraction. The features are selected according to the most informative features using the Chi-Square formula.

3.5 Classification

The selected features will then be used to build a classification model using the naïve Bayes algorithm. This research will use k-fold cross-validation to divide the dataset into training and testing subsets which are carried out repeatedly according to the specified k value.

3.6 Evaluation

Evaluation of model performance will be carried out by applying accuracy, precision, and recall values to the prediction results calculated using the confusion matrix.

4 Result

4.1 Dataset

In this research, the data collected consists of rating attributes and user reviews regarding ChatGPT. The review feature displays ChatGPT user review data in string form. Then the Rating feature displays the data label. A rating of 0 indicates the label is critical thinking and a rating of 1 indicates the label is positive. Some data will be displayed in Table. 1.

Table 1. Dataset reviews ChatGPT

Review	Rating
This is amazing technology. What it needs right now is to give the user controls for the audio listening. It will often cut you off before you're finished talking.	0
Dear openai, This app is gorgeous all around, though I'm experiencing 2 issues, the first being more important, which in my opinion, polish, the app: 1- The speaking system registers on my phone as a call, which, in consequence, forces the software to deteriorate the audio quality.	1
The ChatGPT app is a game-changer in the world of conversational AI. Its ability to understand and generate human-like responses is impressive. The app's interface is user-friendly, making it accessible for both casual users and those seeking more advanced interactions.	1
This app is developed with a great model, and it's a good app to go to if there is anything you need assistance with... I'm impressed with its responses, and the timing is really superfast, so I would recommend this app to everyone before going to almost all the other ones that I have tried so far... great job on getting the app it's a perfect 5-star score, and I hope you receive many more!! 528Hz ðŸ’š	1

I've loved Chat GPT since I first tried it, despite the minor issues it does have (its still early for A.I., there's gonna be issues), but this mobile version actually has a feature where you can have voice conversations with it! It's great!	1
---	---

4.2 Pre-processing Data

After the data is obtained, the data must be cleaned first because the data is still not in good condition to be processed by a computer. After deleting duplicate data and another pre-processing, a total of 27477 rows of data were obtained. Positive data consists of 22642 rows and negative data consists of 4750 rows. From this data, it can be seen that the data is not balanced for each class. Data in the positive class is much more (majority) than data in the negative class. The result of data imbalance is that smaller classes (minority) cannot be predicted. This is because a smaller amount of data makes the machine less able to learn the data and therefore targets the wrong data. To overcome this, the Synthetic Minority Over-sampling Technique (SMOTE) method is used. SMOTE works by increasing the number of samples in the minor class so that the number is the same as the major class by generating new data (synthetic data) based on its nearest neighbors [3]. The pre-processing stages of text data will be shown in Table 2.

Table 2. Pre-processing of ChatGPT user review data

Pre-processing	Input	Output
Case Folding	This is amazing technology. What it needs right now is to give the user controls for the audio listening. It will often cut you off before you're finished talking.	this is amazing technology. what it needs right now is to give the user controls for the audio listening. it will often cut you off before you're finished talking.
Cleaning	this is amazing technology. what it needs right now is to give the user controls for the audio listening. it will often cut you off before you're finished talking.	this is amazing technology what it needs right now is to give the user controls for the audio listening it will often cut you off before you're finished talking
Tokenization	this is amazing technology what it needs right now is to give the user controls for the audio listening it will often cut you off before you're finished talking	this,is,amazing,technology, what,it,needs,right,now,is,t o,give,the,user,controls,for, the,audio,listening,it,will,o ften,cut,you,off,before,you re,finished,talking
Stopwords	this,is,amazing,technology,what, it,needs,right,now,is,to,give,the, user,controls,for,the,audio,listen ing,it,will,often,cut,you,off,befo re,you're,finished,talking,	amazing,technology,needs, right,give,user,controls,audio,listening,often,cut,you're ,finished,talking

Stemming	amazing,technology,needs,right, give,user,controls,audio,listenin g,often,cut,youre,finished,talkin g	amaz,technolog,need,right, give,user,control,audio,list en,often,cut,your,finish,tal k
----------	--	---

4.3 Feature Extraction

At this stage, the data that has gone through pre-processing will have its features extracted by looking at the frequency of occurrence of words in the document. At this stage feature extraction will be carried out using the TF-IDF technique. The results obtained from feature extraction are shown in Fig. 1.

```
tfidf_matrix
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

Fig. 1. Feature extraction results

4.4 Feature Selection

This feature selection uses input namely the results of the TFIDF feature weighting. At the feature selection stage, the best k will be selected using the sklearn library to select a number of the best features. By using several K estimates, the best k is obtained, namely 1500k. A graph showing the accuracy value getting better when the k value approaches 1500 is shown in Fig. 2.

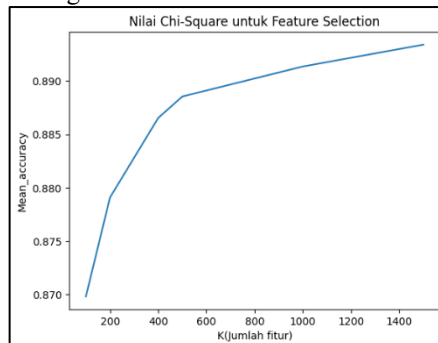


Fig. 2. Chi-Square Value for Feature Selection

4.5 Classification

After the features that have the best contribution are obtained, the next step is to carry out classification. The algorithm used in this research is NB. At the classification stage, k-fold cross validation will be used to provide a more stable and accurate model estimate. The k used in k-fold cross validation is 10.

4.6 Evaluation

The model performance assessment can be seen using the confusion matrix. After the confusion matrix is obtained, other information can be obtained such as accuracy, precision, and recall. The confusion matrix for measuring model performance can be seen in Fig. 3

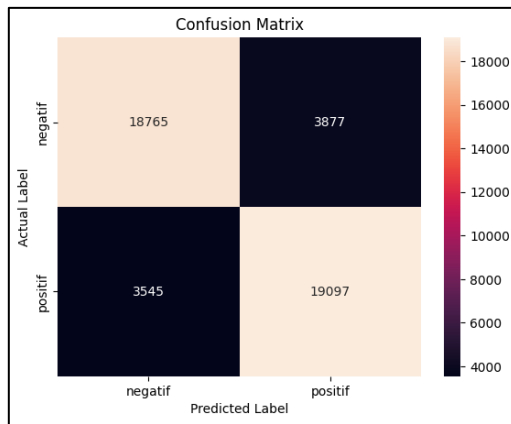


Fig. 3. Confusion Matrix

From the confusion matrix values, accuracy, recall, and precision values can be obtained which are shown in Table 3.

Table 3. Performance Model

Performa	Value
Accuracy	84%
Recall	84%
Precision	83%

4.7 Outsample Data Prediction

After the model data is built, the model will be saved in pickle form for use in the outsample prediction stage. The outsample data then goes through pre-processing and token weighting stages to convert the data into vectors that the computer can understand. Once the data is in vector form, the data is predicted using a model stored in

pickled form. The prediction results for outsample data consisted of 400 data and obtained 208 negative labels and 192 positive labels.

5 Conclusion

From the data obtained in data collection, valid data will then be cleaned to increase classification accuracy. Next, the clean data is weighted using the TF-IDF technique. The weighted data is then selected based on the features that have the most influence on the target feature, using the Chi-Square formula to obtain the best feature with a k value of 1500. Then, using the best features, a model is built using the NB algorithm with a value of $k=10$ for validation cross. The model performance shows 84% accuracy, 84% recall, and 83% precision. Then the model is saved to the dam in pickled form for use in predicting outsample data. From the out-sample prediction data, the results showed that there were 208 negative labels and 192 positive labels.

References

- [1] E. D. Liddy, "SURFACE at Syracuse University Natural Language Processing," 2001.
- [2] B. Liu, "Sentiment Analysis and Mining of Opinions," *Stud. Big Data*, vol. 30, no. May, pp. 503–523, 2012, doi: 10.1007/978-3-319-60435-0_20.
- [3] N. Istiana and A. Mustafiril, "Perbandingan Metode Klasifikasi pada Data dengan Imbalance Class dan Missing Value," *J. Inform.*, vol. 10, no. 2, pp. 101–108, 2023, doi: 10.31294/inf.v10i2.15540.
- [4] I. Rish, "An empirical study of the naive Bayes classifier," *Comput. Sci. Math.*, vol. 3, no. 22, pp. 41–63, 2001, doi: 10.1039/b104835j.
- [5] J. D. M. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," *Proceedings, Twent. Int. Conf. Mach. Learn.*, vol. 2, no. 1973, pp. 616–623, 2003.
- [6] C. H. Yutika, A. Adiwijaya, and S. Al Faraby, "Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 422, 2021, doi: 10.30865/mib.v5i2.2845.
- [7] T. Jain *et al.*, "Sentiment Analysis on COVID-19 Vaccine Tweets using Machine Learning and Deep Learning Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, pp. 32–41, 2023, doi: 10.14569/IJACSA.2023.0140504.
- [8] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 2, pp. 225–231, 2020, doi: 10.1016/j.jksuci.2018.05.010.
- [9] A. R. Atmadja, W. Uriawan, F. Pritisen, D. S. Maylawati, and A. Arbain, "Comparison of Naive Bayes and K-nearest neighbours for online transportation using sentiment analysis in social media," *J. Phys. Conf. Ser.*, vol. 1402, no. 7, 2019, doi: 10.1088/1742-6596/1402/7/077029.
- [10] S. H. A. Samsudin, N. M. Sabri, N. Isa, and U. F. M. Bahrin, "Sentiment

- Analysis on Acceptance of New Normal in COVID-19 Pandemic using Naïve Bayes Algorithm,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 581–588, 2022, doi: 10.14569/IJACSA.2022.0130968.
- [11] I. B. G. Sarasvananda, D. Selivan, M. L. Radhitya, and I. N. T. A. Putra, “Analisis Sentimen Pada Pembelajaran Daring Di Indonesia Melalui Twitter Menggunakan Naïve Bayes Classifier,” *SINTECH (Science Inf. Technol. J.)*, vol. 5, no. 2, pp. 227–233, 2022, doi: 10.31598/sintechjournal.v5i2.1241.
- [12] C. D. Manning, P. Raghavan, and H. Schütze, “An Introduction to Information Retrieval,” no. c, 2009, [Online]. Available: <https://nlp.stanford.edu/IR-book/pdf/00front.pdf>
- [13] R. Shanmugam, “Practical text analytics: maximizing the value of text data,” *J. Stat. Comput. Simul.*, vol. 90, no. 7, pp. 1346–1346, 2020, doi: 10.1080/00949655.2019.1628899.
- [14] N. Nicholas and R. Sutomo, “Comparative Analysis of Sentiment Analysis Using the Support Vector Machine and Naive Bayes Algorithm on Cryptocurrencies,” *J. Multidiscip. Issues*, vol. 1, no. 3, pp. 2–19, 2021, doi: 10.53748/jmis.v1i3.22.
- [15] A. Tabassum and R. R. Patil, “A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing,” *Int. Res. J. Eng. Technol.*, no. June, pp. 4864–4867, 2020, [Online]. Available: www.irjet.net
- [16] A. Ado, N. A. Samsudin, M. M. Deris, A. A. Bichi, and A. Ahmed, “Comparative analysis of integrating multiple filter-based feature selection methods using vector magnitude score on text classification,” *Proc. Int. Conf. Ind. Eng. Oper. Manag.*, pp. 4664–4676, 2021.
- [17] M. Wankhade, A. C. S. Rao, and C. Kulkarni, *A survey on sentiment analysis methods, applications, and challenges*, vol. 55, no. 7. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10144-1.
- [18] I. Wickramasinghe and H. Kalutarage, “Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation,” *Soft Comput.*, vol. 25, no. 3, pp. 2277–2293, 2021, doi: 10.1007/s00500-020-05297-6.
- [19] E. Elsaed, O. Ouda, M. M. Elmogy, A. Atwan, and E. El-Daydamony, “Detecting Fake News in Social Media Using Voting Classifier,” *IEEE Access*, vol. 9, no. MI, pp. 161909–161925, 2021, doi: 10.1109/ACCESS.2021.3132022.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

