Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 11, No. 1, November : 2022

Feature Extraction of Amazon Product Alexa Using Supervised Learning

[1]Ayesha Naureen
Research Scholar
dept. of computer science and Engineering
University of Technology,Jaipur,India
nooriekhan005@gmail.com

[2]Dr. Satish NarayanraoGujar
Professor
dept. of computer science andEngineering
University of Technology,Jaipur,India
satishgujar@gmail.com

Abstract— In the modern world, everyone expresses themselves in few way. In our extremely developed and modernized world, a plethora of societal media platforms as well Android apps, like Amazon, Flipkart, Facebook, WhatsApp, and Twitter, are inundated with views with data. Amazon's website or application is one of the most well-known and widely used platforms. Nearly every excited or communal person inclines to commune his/her opinions by means of comments, as well view as primary source of a sentiments. People's moods are revealed by these words in addition to their sentiments. Writing on these sites is often unstructured, thus it must first be prepared using six preprocessing techniques: bag of words(BOW), TF-IDF, word embedding, as well NLP-base characteristics like count. Here In this study, we examine effects of two attributes, TF-IDF word level as well N-gram, on the sentiment analysis dataset of the Amazon Alexa. Classification algorithms were used in the analysis, and the accuracy, precision, and precall performance characteristics were acknowledged.

Keywords— *Sentiment ,Machine Learning, classification, TF-IDF, N-gram*

I. INTRODUCTION (*HEADING 1*)

Because of the increase in content on communal media sites like Amazon, Twitter, as well Facebook, people are expressing their opinion about product, service, as well government guidelines, amongst other things. Amazon, which has millions of active users monthly, has become a primary basis of comment for government, private organizations, and other service providers. Every day, millions of reviews are written on Amazon, resulting in a massive volume of a unstructured text data. Process of analyzing text data created on Amazon for Amazon application is known as text categorization. One method for obtaining insightful data from customer reviews is sentiment analysis. The Amazon sentiment system is used to categorise reviews into three categories: neutral, positive, and negative. In sentiment analysis, many scholars have presented classification methods[1].

the initial step of text preprocessing by sentiment classification. This stage converts unstructured online content that contains noise so that it can be categorized. This approach will preprocess the text to convert the unstructured, noisy content from the web into format which can be categorized. Tokenization, stop word removal, lower case conversion, stemming, as well number removal are example of preprocessing tasks. Extracting the features is the next stage. TF-IDF, word embeddings, count vectors, and bag of words are examples of text features. NLP stands for natural language processing. Selecting the features comes next, and these are often mutual information, information gain, and chi square. Utilizing ML techniques such as SVM is the last stage.

Some academics have looked into impact of a pre-processing techniques upon sentiment analysis of Amazon reviews. This study will examine the effects of a number of factors, including TF-IDF in addition to bag of words, on sentiment analysis performance. To ascertain which feature is superior, six preprocessing techniques be applied, two types of a features (TF-IDF and BOW) are extract from a text, and classification algorithms are utilized. These comments are unable to convey the emotions of the individuals.

Literature review

The study [3] looks at the impact of preprocessing on tweets that contain a lot of symbols, unidentified words, and abbreviations. After eliminating stop words, URLs, punctuation, and user comments, they used an SVM

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 11, No. 1, November : 2022

classifier to determine the importance of slang terms and spelling correction. In this work, vector representations were used for tasks related to Natural Language Processing [4]. The goal of author was to solve sentiment analysis problems by utilizing word vector representations' efficiency. The three most crucial tasks are recovering sentiment words, identifying sentiment words, and predicting text sentiment. They examined the superiority of vector representation over exclusive text data as well as the quality of vectors across a range of domains. The efficiency of additional vector-based characteristics has also been calculated and confirmed using the representations. They state that they have an F1 score of 85.77 percent and an accuracy for a text SA for APP evaluations of 86.35 percent. The impact of preprocessing on a dataset of movie reviews was examined in paper [5].They tried deleting stop words, negations, non-English letters, stemming, and the prefix 'NOT_', as well as using an SVM classifier. The authors of research [6] looked at 4 datasets: HCR, Sentiment 140 (just three thousand Tweets), Sanders, as well Stanford 1k, plus used Bag of a Words, Lexicon-base feature, and Parts of Speech-base feature. They employ SVM in addition to a mix of several machine learning techniques, including Nave Bayes and Logistic Regression. Old method for a sentiment analysis of small text in this paper [7] lack dependence on emotion words as well modify and simply accumulate sentiment of a sentence is to pursue sentiment of small text; however, they were able to allay difficulties by using sentiment structure as well sentiment computation norm. Proposed approach in research demonstrates how depend parsing deduces sentiment structure using relational migration as well attuned distance, which makes a significant contribute to understanding sentiment of brief text. Dissimilar influence of mapping amongst modifier and emotion word is use to determine sentiment of short text. The findings of their experiments support the efficacy of the technique they presented for addressing the issues through sentiment structure. The authors of study [8] looked at text (e-learning feedback) in Greek plus extract parts of a speech as well text-base feature, as well as evaluating impact of these feature on sentiment classification recital. Joseph D. Prusa used 10 different feature selection strategies and four classifiers in his study [9]. They find that the performance of sentiment analysis is enhanced by the use of a feature selection approach. 3 levels of feature extraction techniques were employed by the authors of study [10]. J48, Nave, and SVM have all been applied. Without using them, the authors of paper [11] looked at a variety of feature selection and classification strategies. The authors of research [12] used a Twitter dataset (total 1000 comments) and used a variety of ML then ensemble approaches (majority voting) classify remarks. They employed twitter-particular features as a classification input to a classifier. Trendy study [13], author used an API to get access to a tweets around Samsung galaxy phones and classified them as favourable, bad, or neutral. They then use LEM2 rough set algorithm to invent decision rule base on Samsung galaxy G5 product review on Samsung site. This will aid business analysts in comprehending products in many dimensions and attributes, as well as the relationships between them.

II. PROPOSED APPROACH

Illustrated in Fig 1, started with the Amazon dataset and applied six pre-processing procedures before extracting feature by means of N grams-IDF technique. The following phase involved using classification techniques as well evaluating parameter.

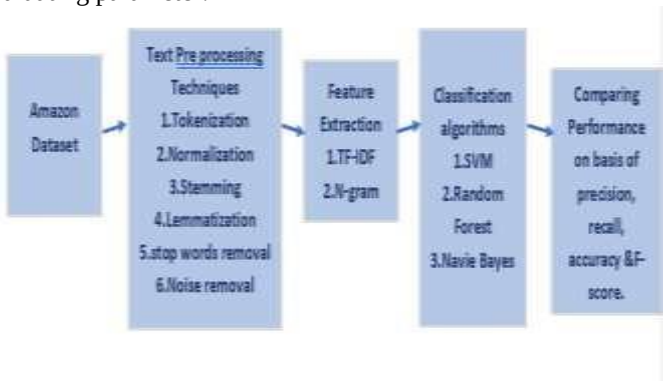


Fig 1: projected methodology

Dataset -Amazon reviews

Amazon reviews for sentimentality strength amazon dataset. Dataset is annotated automatically.

Pre-processign Techniques

Tokenization

Large paragraphs, also known as pieces of text, are divided in this stage keen on tokens, which be essentially sentences. These claims able to further deconstructed into individual words. Examine the following statement that was previously tokenized: Is it not? '?', 'PhD', 'is', 'tough', 'job', 'to', 'do' An arduous task to accomplish after tokenization

Normalization

In order to attain normalization, multiple menu activities are finished concurrently. It includes transforming every text to moreover superior or subordinate case, eliminate punctuation, and translating numerals into words that match to them. enhance uniformity in the way that all texts are prepared.

Stemming

The stemming progression approach is useful in removing unwanted word commutations, such as fishing, fish, fisher to fish, argument, arguing, and arguments towards argue. It is use to modify diverse tenses of a words to their base form.

Lemmatization

Lemmatization is a process of merging 2 or more words into a single word. This looks at word morphology as well eliminates ends like shocked to astonished, caught to caught, and so forth.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 11, No. 1, November : 2022

Eliminating stop Words

The most common terms in English language that are unrelated to sentiment analysis be stop words. Here, stop words that need to be removed are "are," "of," "the," and "at."

Noise Removal

The datasets are not yet processed. We used a combination of regular expressions in natural language processing and manual data cleaning to decrease noises. Noise reduction should be used carefully since it occasionally causes a few rows of data to be lost, which reduces accuracy. The dataset was cleaned using a regular expression that can remove extraneous white space and sort the data into the appropriate column.

Feature Extraction

TF-IDF

The word "frequency" refers to the regularity of an occurrence. One well-known technique for assessing a word's significance within document is inverse document frequency approach. By dividing no.of times a term appears within document by the total number of words in the document (t), the phrase frequency of the particular term (m) is calculated. The relevance of a phrase is ascertained by applying the IDF (Inverse Document Frequency) method. Certain expressions and words, such "is," "an," as well "and," are frequently employed but has no real meaning. IDF (t) = log(N/DF), where DF is a total no.of documents that include the term t and N is the total number of documents. A more effective method of transforming textual data into vector space model to use the TF-IDF format. The term frequency is 10/250=0.04 if a document has 200 words and a mouse appears 10 times in those 200 words; if a document has 50,000 words but only 500 of them contain a mouse, the term frequency is 10/250=0.04. Following that, TF-IDF (mouse) = 0.04*100=4 and IDF (mouse) = 50000/500=100.

N- Gram

N-Gram will create text features for a supervise machine learning methods. n tokens in order from provided text. N may have value of 1, 2, 3, and so forth. We refer to a 'n' as a unigram when its value is 1, a bigram when it is valued at 2, and a trigram when it is valued at 3. An N-gram can be used to identify 'n' consecutive words or sounds in a text or audio sequence. A gramme model is used in a sentiment analysis to evaluate a sentiment of a text or document. We'll use the simplification P to illustrate the probability that the random variable Ui will take on the value "the," or P(Ui = "the"). N words will be denoted by the sequence (the), w1,...w1:n.. Hence, the string z1, z2,..., zn-1 is referenced by w1:n-1. We will use P(z1,z2,z3...zn) to get the joint likelihood of every planet in a sequence with specified value P(U=z1,Y=z2,Z=z3, W=zn). How can we now determine the likelihood of full sequences like P(z1,z2,....zn)? For example, we can use a chain of probabilities to deconstruct this likelihood.

$$P(U1\ldots Un) = P(U1)P(U3|U1:2)\ldots P(Un|U1:n-1)$$

$$\prod_{m=1}^{n} P(Um|U1: m-1) \text{-------- } 1$$

Bi- Gram

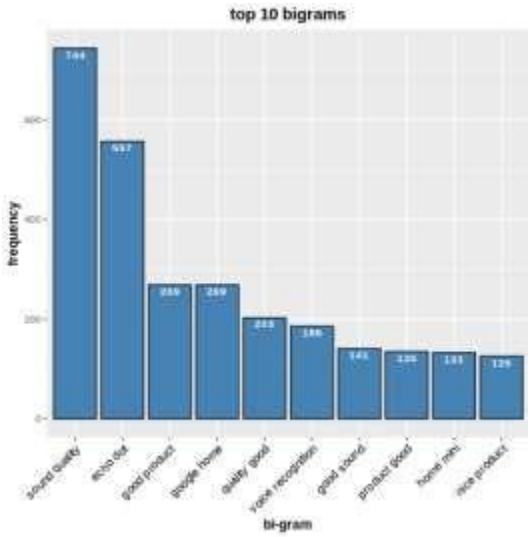
A bigram can be handled as a term in the same manner as individual words in a document. A bigram is a model that uses only the previous word's evaluation to predict the next word; n-2 is the number of words in the previous word. To estimate the likelihood of a word given all preceding words, the Bigram model simply considers the conditional probability of one preceding word. Put otherwise, you approximate it using probability: (that | the) P(that|the) P(that|the) P

Therefore, the following approximation should be used when using the Bigram Model to forecast a conditional probability of a next word:

$P(z_n|z_1^{n-1}) \sim P(z_n|z_{n-1})$

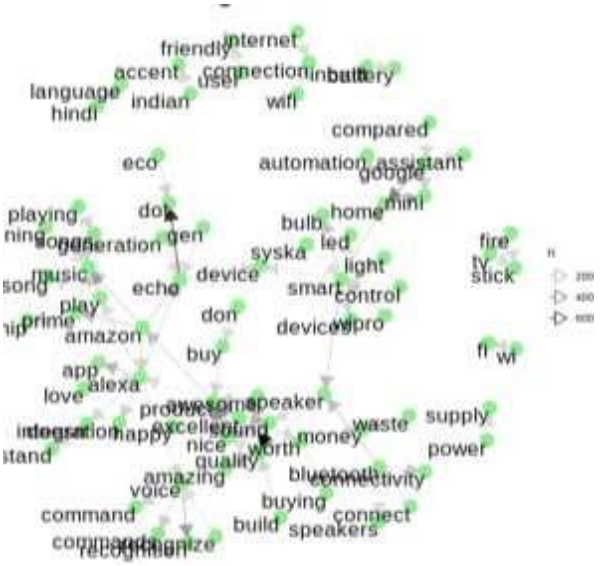
To generate two grams, the value of n=2 is provided to the n-grams function. But before we can provide the text to the n-grams function, we need to divide it up into tokens.

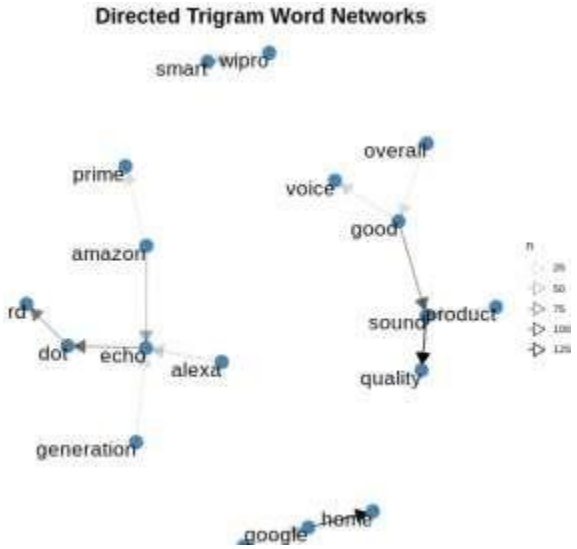
P(zn|zn-1) = P(zn-1,zn) / P(zn-1)[51]. Probability of bigram P(zn|zn-1) = P(zn-1,zn) / P(zn-1)



Several terms, such as sound, music, voice, and google, appear in several bigrams in various contexts.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 51, Issue 11, No. 1, November : 2022

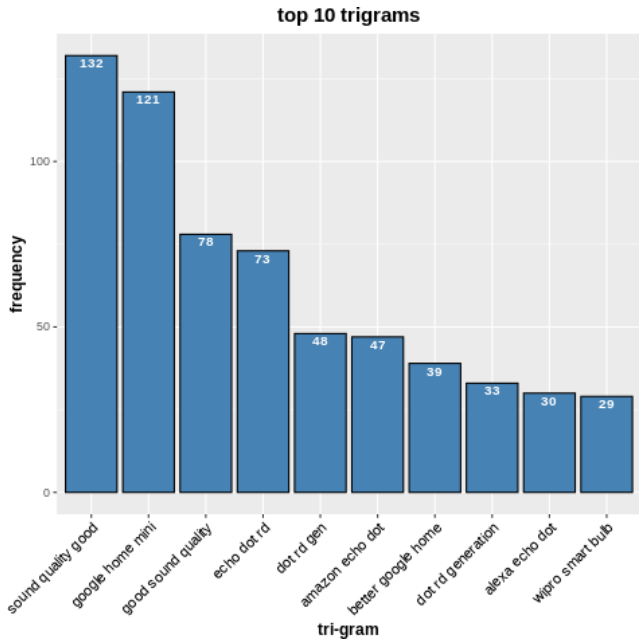




Below fig is of word cloud of tri grams

Word Cloud importance of sound quality and the comparison to a Google product are two important themes. Product and Amazon service satisfaction are also prevalent.Tri- gram

It's a trigram model if two previous words are taken into account. Trigrams are a subset of n grammes when n equals three.



Although the terms 'echo' and 'dot' exist in several trigrams, they might be regarded stop words in this case. Sound is the most common word in frequent trigrams indicating positive consumer feedback on sound quality.

III. CLASSIFICATION ALGORITHMS

SVM

This is a useful approach for mutually regression as well classification. Toward distinguish classes, draws hyperplane. It approach perform exceptionally well through regression, and influence of a SVM grows as number of dimensions growths. When dimension numeral is extra than sample number, SVM perform well [15]. There is a disadvantage as well: it does'nt perform well on huge datasets. To improve its computing efficacy, SVM makes significant use of cross-validation.

Navie Bayes

This is a sophisticated classification approach that uses probability to classify data. With millions of records, this algorithm also performs admirably. It's base on Bayes theory as well classifies data usage of innumerable probabilities. The class with a maximum probability is predicted class in the Nave Bayes algorithm. Nave Bayes is

also well-known as Maximum Posterior Bayes. Nave Bayes offers a variety of benefits and drawbacks in several fields. It's a nimble and scalable algorithm. Mutually Multiclass as well Binary Classification been used. It might be use on tiny datasets plus, then it produces upright results[14].

Random Forest

It's a collection of decision tree methods that can be used to classify and predict data. More trees, in general, resemble to improved performance as well competence in this method. Using the bootstrap approach, excerpt example set of a data opinions from specified training set. Create decision tree base on results of the earlier phase. We will receive the number of trees if we apply the previous two procedures (in case 100). Every tree that is built will cast a vote meant for data opinion. Calculate decision tree classifier's majority voting [16].

IV. PERFORMANCE PARAMETERS

Precision: It indicates how precise the classifier is. It's the proportion of successfully anticipated positive reviews to total no. of positive reviews expected.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (1)$$

Recall: It assesses the classifier's completeness. It's the proportion of precisely predicts positive reviews in the corpus to the actual no. of positive reviews in corpus.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (2)$$

F-measure: Precision and recall have a harmonic mean. The best value for an F-measure is 1 plus lowest value be 0. F-measure could calculated using formula [58].

$$\text{Measure} = \frac{z-Precisions*Recall}{Precision+Recall} \qquad (3)$$

Accuracy: It's measured as ratio of a number of appropriately predict reviews to total no.of reviews in corpus, and it is one of the most popular performance evaluation parameters. The following is the formula for calculating accuracy:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

V. RESULTS AND DISCUSSIONS

Two characteristics of the Amazon sentiment analysis dataset were examined: 1. Word level (TF-IDF) and 2. Amazon reviews (N-Grams). Table 1 presents the outcomes of three classification algorithms (SVM, Random Forest, and Naive Bayes) that make use of the TF-IDF feature (four performance parameters: f-score, accuracy, precision, as well recall). Table 2 presents outcomes of four performance criteria (accuracy, precision, recall, as well f-score)-based classification methods (Random Forest, Naive Bayes, SVM, and N-gram features). As seen from a two tables, SVM functions effectively in both situations. Figure 2 illustrates this goal of determining which features perform better than others.

Amazon Dataset (Word Level TF-IDF)				
ML Algorithm	Accuracy%	Precision%	Recall%	F-Score%
SVM	54	56	50	50
NB	53	52	44	42
RF	51	47	44	44

Table 1: As can be observed in fig., the feature performs 3–4% better at the word level since it regards every word as equally important.

Amazon Reviews Dataset (N – Grams Vectors)				
ML algorithms	Accuracy%	Precision%	Recall%	Score%
SVM	51	49	43	42
NB	50	41	41	38
RF	50	44	39	41

Table 2: N gramme feature on Amazon reviews.

CONCLUSION

Using the Amazon reviews dataset, three different classification algorithms were employed in this study, accounting for two characteristics (TF-IDF as well N-Grams). later than conducting sentiment analysis on evaluations, we found TF-IDF feature perform 3–4% better than a N-Gram features. Therefore, we may infer that TF-IDF is a better feature alternative than N-Gram when utilizing machine learning algorithms for text classification. By delivering maximum output for every 4 comparison metrics—accuracy, recall, precision, as well f-score—as well as for mutually feature extraction techniques—N-Gram and word-level TF-IDF—we found that SVM produced the best sentiment predictions. SVM is therefore the best sentiment analysis technique, as well feature extraction technique taken as a whole are suitable In the future, more variables like word embeddings, Amazon-only properties, word polarity score features, and so forth will be compared.

REFERENCES

[1] Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." In Proceedings of the workshop on languages in social media, pp. 30-38. Association for Computational Linguistics, 2011

[2] Mohammad, Saif M., Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. "Sentiment, emotion, purpose, and style in electoral tweets." Information Processing & Management 51, no. 4 (2015): 480-499

[3] Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." In Proceedings of the workshop on languages in social media, pp. 30-38. Association for Computational Linguistics, 2011

[4] X. Fan, X. Li, F. Du, X. Li, and M. Wei, "Apply word vectors for sentiment analysis of APP reviews," 2016 3rd International Conference on Systems and Informatics(ICSAI), Shanghai,2016,pp.1062-1066

[5] Shi, Y., Xi, Y., Wolcott, P., Tian, Y., Li, J., Berg, D., Chen, Z., Herrera-Viedma, E., Kou, G., Lee, H., Peng, Y., Yu, L. (eds.):

Proceedings of the First International Conference on Information Technology and Quantitative Management, ITQM

[6]] Fouad M.M., Gharib T.F., Mashat A.S. (2018) Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble. In: Hassanien A., Tolba M., Elhoseny M., Mostafa M. (eds) The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018). AMLTA 2018. Advances in Intelligent Systems and Computing, vol 723. Springer, Cham

[7] J. Li and L. Qiu, "A Sentiment Analysis Method of Short Texts in Microblog," 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou,2017,pp.776-779

[8] Spatiotis N., Paraskevas M., Perikos I., Mporas I. (2017) Examining the Impact of Feature Selection on Sentiment Analysis for the Greek Language. In: Karpov A., Potapova R., Mporas I. (eds) Speech and Computer. SPECOM 2017. Lecture Notes in Computer Science, vol 10458. Springer, Cham

[9] Prusa, Joseph D., Taghi M. Khoshgoftaar, and David J. Dittman. "Impact of Feature Selection Techniques for Tweet Sentiment Classification " In FLAIRS Conference, pp. 299-304. 2015.

[10] Angulakshmi, G., and Dr R. Manicka Chezian. "Three level feature extraction for sentiment classification." International Journal of Innovative Research in Computer and Communication Engineering

[11] Kamale, Mr Amit S., Pradip K. Deshmukh, and Prakash B. Dhainje. "A Survey on Classification Techniques for Feature-Sentiment Analysis." International Journal on Recent and Innovation Trends in Computing and Communication 3, no. 7 (2015): 4823-4829.

[12]] Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, Prof. Dr. D. R. Ingle “Sentiment Analysis in Twitter” International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 01 | Jan-2018 pages 880-886.

[13] Das, Tushar Kanti, D. P. Acharjya, and M. R. Patra. "Opinion mining about a product by analyzing public tweets in Twitter." In Computer Communication and Informatics (ICCCI)

[14] M. Rathi, A. Malik, D. Varshney, R. Sharma, and S. Mendiratta, "Sentiment Analysis of Tweets Using Machine Learning Approach," 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, 2018, pp. 1-3

[15] İ. İşeri, Ö. F. Atasoy and H. Alçiçek, "Sentiment classification of social media data for telecommunication companies in Turkey," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 1015-1019

[16]] Breiman, L., Random forests. Mach. Learn. 45(1):5–32