Advanced Engineering Days

aed.mersin.edu.tr

Accuracy prediction of identification in remote customer acquisition in banking with machine learning

Hazim Iscan[1], Seyma Nur Alkan*[2]

[1]Konya Technical University, Faculty of Engineering and Natural Sci, Computer Engineering, Türkiye, hiscan@ktun.edu.tr
[2]Konya Technical University, Graduate Education Institute, Computer Engineering, Türkiye, alkanseymanur@gmail.com

Cite this study: Iscan, H., & Alkan, S.N. (2022). Accuracy prediction of identification in remote customer acquisition in banking with machine learning. 2nd Advanced Engineering Days, 76-78

Keywords
Machine learning
Natural language processing
Word2Vec
Word embedding

Abstract

In banking, thanks to remote identity detection, the customer representative and the real person will not need to be physically in the same environment, and new customers will be reached quickly and effectively. In this study, a data set consisting of randomly generated contact information for remote identity detection methods that can be used in customer identity verification is trained using Natural Language Processing Techniques and an estimate is made as to whether the person is real or not. One of the methods used in this study is "Word Embedding". Word Embedding is a method for closely representing words with similar meanings. The generated data set is modeled with Word2Vec, a word vector algorithm. The clustering of the word vectors obtained by Word2Vec techniques, in terms of their formal properties as well as the semantic relations of the words they belong to, has been examined. Two different Word2Vec methods such as CBoW and Skip-Gram were used to create the model. According to the results of the application, a success rate of 89% was achieved in the estimation of the correct data.

Introduction

The Banking Regulation and Supervision Agency of Turkey, BRSA, published the "Regulation on Remote Identification Methods used by Banks and Establishment of Contractual Relationship in Electronic Environment" about remote authentication methods that will be used by banks to remotely gain new customers and verify customer identity on 01 April 2021.

When the global approaches in distance identity detection processes are examined, it is observed that digital methods predominantly come to the fore. The increase in digitalization and the use of artificial intelligence technologies have increased the studies on this subject. Bektaş et al. using the distinguishing features of the characters, tried to identify the characters with the best accuracy rate of a document in picture format with the help of classification methods. By comparing the performance results and duration of the classification methods, they determined the best method among them [2]. Llados et al. describe the ICAR system, an application for automatic reading of identity cards and passports. The type and content of the document are recognized by a number of complementary statistical and structural OCR techniques. Although the system was originally designed for Spanish documents, it allows the integration of new formats through a supervised learning procedure [3]. Dwi et al. compared the size of original, medium and small images in color and grayscale images using Optical Character Recognition (OCR) technology for ID card reading. As a result of the comparison, they found the accuracy rate of grayscale data to be 88.58% and color data to be 86.32% [4].

Material and Method

Natural Language Processing

NLP, or Natural Language Processing, aims to understand or reproduce the canonical structure of natural languages by analyzing them. The convenience that this analysis will bring to people can be summarized with many topics such as automatic translation of written documents, question-answer machines, automatic speech and command comprehension, speech synthesis, speech generation, automatic text summarization, and information provision. The widespread use of computer technology has enabled specialist software produced from these titles to enter every area of our daily life [5].

Gensim Python Library

Gensim is an open source python library for natural language processing and was developed by Czech natural language processing researcher Radim Řehůřek. Gensim library enables to parse words by training Word2Vec models on a special corpus, with CBOW or Skip-Gram algorithms. [6,7].

Word2Vec

Word2vec is a two-layer neural network that processes text by "vectoring". Its input is a text string and its output is a set of vectors. The word vectors method is proposed to represent the words as vectors in an n-dimensional space and to determine the semantic similarity between them by calculating the distances between the words in this way. Word2Vec is a neural network based approach for embedding words. Trained with a large text set, this model generates a unique vector for each word in high-dimensional space. The characteristic of these unique vectors created is that words with similar meanings in the dataset form vectors close to each other. It has two methods, CBoW and Skip-gram [7,8].

CBoW uses a context surrounding the word to predict a word, while Skip-gram tries to guess the word by surrounding words with a fixed window size. Skip-Gram can produce better results for sparse words. CBoW and Skip-Gram methods are shown in Fig. 1 [9].

As a result of the Word2Vec operation, a dictionary is obtained in which each word has a vector. Since the classification of person information is done in this study, the aim is to extract the vectors of each attribute using this data. In order to achieve this, firstly, each word in the data set is matched with its vectors and a matrix of size k x w is obtained. Here k represents the number of words in the document and w represents the vector size.

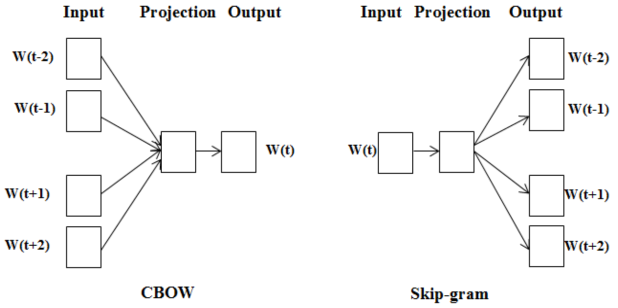


Figure 1. Word2Vec (CBow and Skip-gram) [8]

Skip-gram

In the skip-gram, the input is the target word, while the outputs are the words around the target word. It is aimed to represent the inputs and outputs in the most appropriate way semantically, by comparing them probabilistically with each other.

CBoW (Continuous Bag of Words)

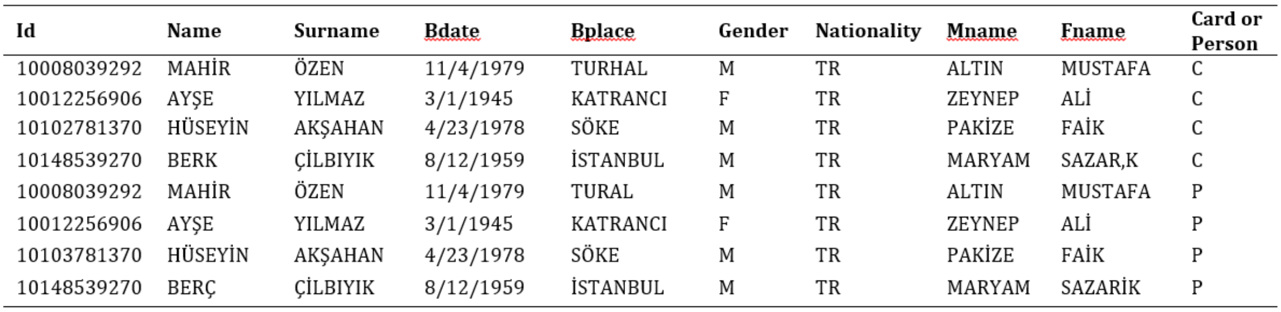
CBoW (Continuous Bag of Words) is a very similar approach to Skip-Gram. The only difference is that the inputs and outputs are swapped. The idea is that when the words around a word are given to the system, it wants to know which word is most likely to appear in those words.

Application

This study was carried out using the Python programming language and machine learning libraries prepared in that language. The dataset was randomly generated for this study. It is not related to real persons or institutions.

An example of the dataset is shown in Table 1. Personal information is classified as 'P' and 'C' in the 'Card Or Person' attribute. 'P' is the information from the form that the person has filled in, and 'C' is the information from the person's Identity Card. The information received from the Identity Card is correct information. There are various errors in the information entered by the person.

Table 1. Example dataset



Id	Name	Surname	Bdate	Bplace	Gender	Nationality	Mname	Fname	Card or Person
10008039292	MAHİR	ÖZEN	11/4/1979	TURHAL	M	TR	ALTIN	MUSTAFA	C
10012256906	AYŞE	YILMAZ	3/1/1945	KATRANCI	F	TR	ZEYNEP	ALİ	C
10102781370	HÜSEYİN	AKŞAHAN	4/23/1978	SÖKE	M	TR	PAKİZE	FAİK	C
10148539270	BERK	ÇİLBIYIK	8/12/1959	İSTANBUL	M	TR	MARYAM	SAZAR,K	C
10008039292	MAHİR	ÖZEN	11/4/1979	TURAL	M	TR	ALTIN	MUSTAFA	P
10012256906	AYŞE	YILMAZ	3/1/1945	KATRANCI	F	TR	ZEYNEP	ALİ	P
10103781370	HÜSEYİN	AKŞAHAN	4/23/1978	SÖKE	M	TR	PAKİZE	FAİK	P
10148539270	BERÇ	ÇİLBIYIK	8/12/1959	İSTANBUL	M	TR	MARYAM	SAZARİK	P

As data preprocessing, the text attributes to be used in the model are converted to lowercase with the lower() method. With the preprocessed data set, two different models were created using CBoW and Skip-Gram methods. After the models were created, the similarity between the two models was calculated with the Word2vec model.similarity( ) function.

Results

According to the results of the application, it is not sufficient to select the model only according to the accuracy values. According to the results of the analysis, CBoW and Skip-Gram algorithm; found the correct data 89% correct and the incorrect data -09% incorrect, and made a correct prediction.

CBoW models generally work better on smaller datasets, while Skip-gram models work better on larger datasets. CBoW requires less computing power, while Skip-Gram requires more computing power. Skip-gram gives better results while CBoW is not good at understanding two or more meaningful words.

Conclusion

For the same data set, the classification process was carried out successfully by evaluating the CBoW and Skip-Gram algorithms separately. Although the classification success rates of the methods differ, it has been seen that each algorithm is sufficient in classification.

References

[1] BDDK. Uzaktan Kimlik Tespiti Yöntemlerine ilişkin Yönetmelik. [Online] Available: https://www.resmigazete.gov.tr/eskiler/2021/04/20210401-7.htm, Accessed on: Feb. 26, 2022

[2] Bektaş, B., Babur, S., Turhal, U., & Köse, E. (2016, Kasım). Makine Öğrenmesi Yardımıyla Optik Karakter Tanıma Sistemi. 5. Uluslararası Matbaa Teknolojileri Sempozyumu, 487-494, İstanbul, Türkiye

[3] Lladós, J., Lumbreras, F., Chapaprieta, V., & Queralt, J. (2001). ICAR: Identity Card Automatic Reader. 470-474. 10.1109/ICDAR.2001.953834.

[4] Dwi H., Purnomo H., & Purwanto H. (2019). Utilization of Optical Character Recognition Technology in Reading Identity Cards. Int. Journal of Information Technology and Business, Vol. 2, No. 1, 38-46

[5] Simsek, H.K. Makine Öğrenmesi Dersleri 6: NLP'ye Giriş. [Online] Available: https://medium.com/data-science-tr/makine-%C3%B6%C4%9Frenmesi-dersleri-6-do%C4%9Fal-dil-i%CC%87%C5%9Fleme-nlp-453c3c6b062a, Accessed on: Feb. 22, 2022

[6] Python website. [Online] Available: https://www.python, Accessed on: Feb. 25, 2022

[7] Li, Z. A Beginner's Guide to Word Embedding with Gensim Word2Vec Model [Online] Available: https://towardsdatascience.com/a-beginners-guide-to-word-embedding-with-gensim-word2vec-model-5970fa56cc92, Accessed on: Feb. 20, 2022

[8] Aravind, C. R. (2022). Word Embeddings in NLP | Word2Vec | GloVe | fastText. [Online] Available: https://medium.com/analytics-vidhya/word-embeddings-in-nlp-word2vec-glove-fasttext-24d4d4286a73, Accessed on: Feb. 20, 2022

[9] Mikolov, T., Le, Q., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. Available: https://arxiv.org/pdf/1309.4168v1.pdf, Accessed on: Feb. 21, 2022