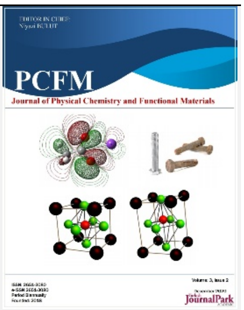
Journal of Physical Chemistry and Functional Materials

Home Page of Journal: https://dergipark.org.tr/jphcfum



Extracting Meaningful Information from Turkish Chemistry and Physics Texts with Machine Learning

Mucahit KARADUMAN*[,a], Muhammed YILDIRIM[b]

[a]*Department of Software Engineering, Malatya Turgut Ozal University, Malatya, Turkey*
[b]*Department of Computer Engineering, Malatya Turgut Ozal University, Malatya, Turkey*
* *Corresponding author: E-mail: mucahit.karaduman@ozal.edu.tr*

ABSTRACT

This study emphasizes the importance of processing a dataset consisting of Turkish chemistry and physics texts created by us through artificial intelligence systems. A model is proposed to pave the way for artificial intelligence-based analyses and discoveries in the basic sciences of chemistry and physics. Chemistry and physics, the basic sciences, are critical in many industrial, medical, and environmental applications. However, significant data analysis is required to access and understand information in these areas. This study aims to demonstrate the effectiveness of machine learning methods in extracting meaningful information from Turkish chemistry and physics texts. For this purpose, the tokenization process is first performed, and then the features are extracted with Term frequency-inverse document frequency (TF-IDF) and Bag-of-Words (BOW) methods. The combined features are classified separately with Support Vector Machine (SVM), Naive Bayes (NB), Quadratic Discriminant Analysis (QDA), k-nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB) algorithms. According to the classification results, the best calculation time and the most successful accuracy rate are obtained with NB at 95%. These results are essential for artificial intelligence systems to understand and process information correctly. It shows that scientists and researchers can access information faster and accelerate scientific discovery using Turkish sources. Such artificial intelligence models can also be essential in education, providing students with a more effective and personalized learning experience. Therefore, processing Turkish chemistry and physics texts with artificial intelligence systems is essential in including studies conducted in this language in global studies in scientific research, education, and industrial applications.

ARTICLE INFO

Keywords:
Artificial Intelligence,
Bag-of-Words,
Classification,
Machine Learning,
TF-IDF

Received: 2024-06-15
Accepted: 2024-07-26
ISSN: 2651-3080
DOI: 10.54565/jphcfum.1501651

1. INTRODUCTION

Nowadays, large amounts of data are collected and stored in scientific research and industrial applications. Especially in primary science fields such as chemistry and physics, the analysis and interpretation of these data form the basis of discoveries and advances. However, much of this data may be in different languages and formats, complicating analysis processes. Artificial intelligence techniques are essential in making this complex and big data meaningful.

Scientific discoveries and technological advances are closely related to analyzing and making sense of large amounts of data. However, most of this data may be in different languages and formats, which can pose a significant obstacle for scientists. Making sense of the information obtained from Turkish sources is essential for researchers working in this language and artificial intelligence systems that use this information.

Figure 1. Word cloud in the fields of Chemistry and Physics

In particular, an in-depth examination of Turkish sources in primary science fields such as chemistry and physics will allow local scientific communities to strengthen and reach a wider audience. Therefore, processing Turkish scientific texts with artificial intelligence systems will contribute to scientific discovery and enable the dissemination of Turkish science to a broader audience.

Turkish researchers need to analyze scientific content in Turkish chemistry and physics texts in depth. However, analyzing and making sense of these texts often encounters language and information processing challenges. In particular, artificial intelligence models and deep learning techniques in this field are generally focused on English sources and are not sufficiently developed to work with Turkish texts.

This may lead to incomplete or misunderstanding of Turkish scientific literature, hinder the effective sharing of scientific knowledge, and harm the development of local scientific communities. Therefore, processing Turkish chemistry and physics texts with artificial intelligence systems and extracting meaningful information will enable scientists and researchers to benefit more from Turkish sources and accelerate scientific discovery.

This study aims to develop an artificial intelligence-based model using a data set consisting of Turkish chemistry and physics texts and evaluate its accuracy. This model will facilitate the analysis and interpretation of Turkish scientific literature, and artificial intelligence systems will understand the scientific studies of Turkish researchers, enabling their participation in science to be more effective.

This study aims to develop an artificial intelligence-based model using a data set consisting of Turkish chemistry and physics texts. First of all, we will clean and tokenize the texts by passing the data set we obtained through pre-processing steps. These steps are essential for understanding the texts correctly and training the model effectively.

Then, the feature extraction phase will begin. At this stage, features are extracted using two different methods, and these features are combined. After this process, as the last stage, the classification performances of the data are examined and compared with the classification algorithms.

Precision, recall, and F1 scores are compared for the accuracy rates of our model. Basic sciences such as chemistry and physics form the basis of many scientific developments. Many studies have been conducted using artificial intelligence in these two areas. In their research, Debus et al. [1] found that deep learning tools are used in analytical chemistry to extract qualitative and quantitative information from high-dimensional and complex chemical measurements. Fooshee and his colleagues [2] conducted the training and testing phases on the dataset they prepared using MLP and LSTM, which are deep-learning methods for predicting chemical reactions and reaction pathways. As a result of tahini, they obtained an accuracy rate of 99.0% with MLP and 94.6% with LSTM. Again, in the research conducted by Cova et al.[3], they stated that deep learning applications are at the beginner level in chemistry. They especially mentioned the need for further progress in interpreting the results. Goh et al. [4], based on the results of their recent study, have shown that deep learning is used in fields such as computer-aided drug design, computational structural biology, quantum chemistry, and material design, which are chemical studies on computation, and that it produces more successful results than traditional classification methods. Rajan et al. [5] studied detecting chemical strings from the Chemical image dataset and obtained results using deep learning. Jha et al. [6] wanted to predict material properties accurately and quickly with the deep neural network model ElemNet they designed. They also identified physical and chemical interactions and similarities for different elements with the model they developed. Artificial intelligence models are also used in the drug development phase related to chemistry [7], [8]. Studies are using artificial intelligence in the field of physics. An artificial intelligence model has been proposed to predict issues such as quantum measurement, analysis, and parameter estimation [9]. One of the machine learning models that combines techniques and processes in heterogeneous scientific fields and is used to obtain results in these fields is the Symbolic regression (SR) method [10].

The proposed method is included in the second part of this proposed study on the classification of texts in the fields of chemistry and physics. Information on designing the method and other calculations is included in this section. In the third section, the application of the model to the text dataset and its performance values are given. In the last section, the contribution of the proposed model to science and future studies is evaluated.

2. MATERIAL AND METHOD

Figure 2 shows the proposed model for analyzing basic science texts, physics, and chemistry. The model starts by creating the Physics and Chemistry text dataset and applying pre-processing to it. Then, the features extracted in the model continue with feature extraction, and the classification stage is reached. At this stage, it is decided which scientific field the text entered for the test belongs to.

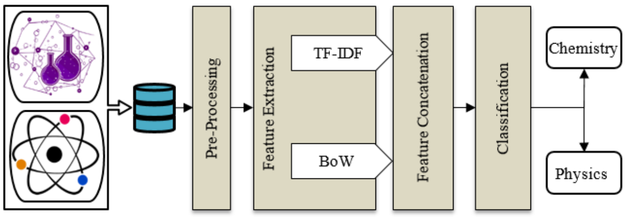


Figure 2. Proposed method

2.1. Dataset

A dataset was created for this study to investigate the primary science fields of Physics and Chemistry. Sentences were determined to include the sub-topics of these two scientific fields. The sub-fields to which the sentences determined for chemistry belong are atoms and periodic systems, chemical interactions, matter, mixtures, acids and bases, atoms, solutions, energy, electricity, and reactions. For physics, topics such as matter and its properties, motion and force, energy, heat and temperature, electrostatics, magnetism, pressure, buoyancy, waves, optics, atomic physics, radioactivity, simple machines, gravity, momentum, shots, and quantum are included. A dataset consisting of 63,570 words was created, bringing together the topics and contents in a balanced manner in these two science fields.

2.2. Tokenization Process

Tokenization is dividing texts into small pieces called tokens in natural language processing. The basic principle of dividing a text document into tokens is the division of tokens, such as spaces and punctuation marks [11]. In the tokenization process, first of all, unnecessary characters and spaces are removed from the text and converted to lowercase. The text is divided into words and sentences using markers. Each piece formed becomes a token. Stored tokens are used in text analysis in subsequent stages.

2.3. Feature Extraction

Term frequency-inverse document frequency (TF-IDF) is a text feature extraction technique widely used in text mining and information retrieval fields. TF-IDF is used to determine how important a term is in a document. This feature considers the frequency (TF) of a term within the document and how common that term is in other documents (IDF). TF is calculated as the ratio of a term to the total number of words in a document. The more frequently a term is used, the more critical it is. Inverse Definitiveness of a Term (IDF) measures how common a term is in a document relative to others. If a term is rarely used generally, it is considered more descriptive and has a higher IDF value. This indicates that that term better describes the document. By multiplying these two values, the TF-IDF value is calculated and kept in a matrix [12].

Bag-of-Words (BoW) counts words and represents them as a vector for the document. A vector is created for each document with CountVectorizer. While creating these vectors, a score is obtained for each word by calculating the frequency of each word in the document. The word's importance is emphasized by ranking the score value obtained [12].

A new feature vector is obtained by analyzing the texts using two different methods and combining the features extracted from the calculations for the words. This newly created vector represents all the text content used. After this stage, operations are performed using this new vector.

2.4. Classification Methods

Seven different classifiers were used to classify the features obtained using the dataset from the developed model. At this stage, Support Vector Machine (SVM) [13], Naive Bayes (NB) [14], Quadratic Discriminant Analysis (QDA) [15], k-nearest Neighbors (KNN) [16], Logistic Regression (LR) [17], Random Forest Classifier (RF) [18] and Gradient Boosting Classifier (GB) [19] are used. These are the classifiers used in the classification process.

3. EXPERIMENTAL RESULTS

The prepared dataset consists of sentences containing topics in chemistry and physics. The subject of each sentence is determined by labeling which basic science it is. At this stage, the dataset is pre-processed by performing tokenization. The dataset's unnecessary spaces and punctuation marks are removed, and the text is converted to lowercase.

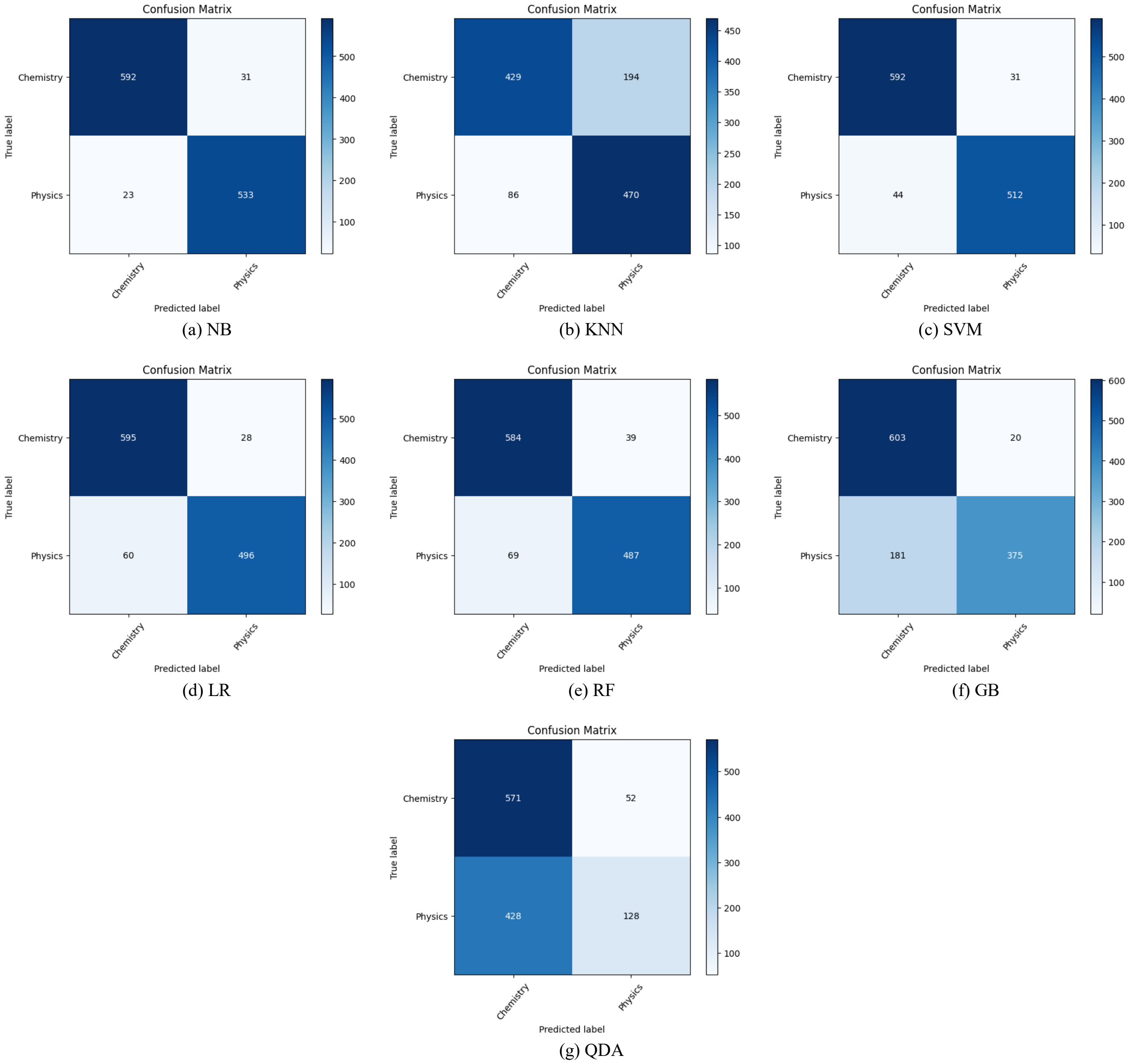
Then, the value of individual words is calculated using TF-IDF and BoW methods, and a new hybrid vector is obtained by combining the resulting feature vectors. This hybrid vector contains the values obtained from TF-IDF and BoW operations for each tokenized word.

The final stage is the classification of this calculated hybrid vector. In the classification phase, the information in the original text, divided into training and testing, is used. While 80% of the information in the dataset is used for training, the remaining 20% is used for testing. Seven different algorithms are used for classification, and the results are compared. The algorithms used are SVM, NB, QDA, KNN, LR, RF, and GB.

Table 1. Performance comparison of classification methods

Methods	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Support**	**Time(ms)**
SVM	94	94	94	1179	6036.54
NB	95	95	95	1179	424.15
QDA	64	59	54	1179	149113.00
KNN	77	76	76	1179	4290.28
LR	93	93	93	1179	1000.68
RF	90	90	90	1179	5511.48
GB	85	83	82	1179	4362.70

Table 1 gives the precision, recall, and F1-score obtained as a result of the classification. When Table 1 is examined, the NB method achieves the best performance with a rate of 95%.



(a) NB (b) KNN (c) SVM

(d) LR (e) RF (f) GB

(g) QDA

Figure 3. Confusion matrices for classification results

Confusion matrices were calculated for the classification results. The confusion matrix for NB, where the best results were obtained, is given in Figure 3(a). The result that should have been 623 for chemistry was estimated as 615, while the result that should have been 556 for physics was calculated as 564. The calculation time for NB is 424.15ms, and the result is obtained with the best calculation time. When the confusion matrices given for other classification algorithms are examined, it is seen that SVM ranks second in terms of success rate. It is seen that the worst result was obtained with QDA with 64%. At the same time, it is seen that QDA gives the worst result regarding calculation time. Figure 3 shows the confusion matrices of all other methods.

4. CONCLUSIONS

Since chemistry and physics are essential basic sciences for developing many branches of science, this study aims to recognize and correctly classify texts for artificial intelligence systems to analyze studies in this field. Thus, synthesizing new and old information in this field and extracting information can be achieved more effectively. The text dataset created for this purpose was analyzed and classified. 7 different classification algorithms were used for this classification, and the best result was obtained with NB. At this point, the ability to distinguish chemistry and physics texts from each other with artificial intelligence systems is achieved with a 95% accuracy rate. Among the studies that can be done in the later stages, the sub-branches of these basic sciences may be added to the dataset separately, and their classifications may be made.

Competing interests

The authors declare that they have no competing interests.

REFERENCES

[1] B. Debus, H. Parastar, P. Harrington, and D. Kirsanov, "Deep learning in analytical chemistry," *TrAC - Trends in Analytical Chemistry*, vol. 145, p. 116459, 2021, doi: 10.1016/j.trac.2021.116459.

[2] D. Fooshee *et al.*, "Deep learning for chemical reaction prediction," *Molecular Systems Design and Engineering*, vol. 3, no. 3, pp. 442–452, 2018, doi: 10.1039/c7me00107j.

[3] T. F. G. G. Cova and A. A. C. C. Pais, "Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns," *Frontiers in Chemistry*, vol. 7, no. November, pp. 1–22, 2019, doi: 10.3389/fchem.2019.00809.

[4] G. B. Goh, N. O. Hodas, and A. Vishnu, "Deep learning for computational chemistry," *Journal of Computational Chemistry*, vol. 38, no. 16, pp. 1291–1307, 2017, doi: 10.1002/jcc.24764.

[5] K. Rajan, A. Zielesny, and C. Steinbeck, "DECIMER: towards deep learning for chemical image recognition," *Journal of Cheminformatics*, vol. 12, no. 1, pp. 1–9, 2020, doi: 10.1186/s13321-020-00469-w.

[6] D. Jha *et al.*, "ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition," *Scientific Reports*, vol. 8, no. 1, pp. 1–13, 2018, doi: 10.1038/s41598-018-35934-y.

[7] C. Hasselgren and T. I. Oprea, "Artificial Intelligence for Drug Discovery: Are We There Yet?," *Annual Review of Pharmacology and Toxicology*, vol. 64, pp. 527–550, Jan. 2024, doi: 10.1146/ANNUREV-PHARMTOX-040323-040828.

[8] J. Zhang *et al.*, "Artificial Intelligence Enhanced Molecular Simulations," *Journal of Chemical Theory and Computation*, vol. 19, no. 14, pp. 4338–4350, Jul. 2023, doi: 10.1021/ACS.JCTC.3C00214.

[9] M. Krenn, J. Landgraf, T. Foesel, F. M.-P. R. A, and undefined 2023, "Artificial intelligence and machine learning for quantum technologies," *APS*, vol. 107, no. 1, Jan. 2023, doi: 10.1103/PhysRevA.107.010101.

[10] D. Angelis, F. Sofos, and T. E. Karakasidis, "Artificial Intelligence in Physical Sciences: Symbolic Regression Trends and Perspectives," *Archives of Computational Methods in Engineering*, vol. 30, no. 6, pp. 3845–3865, Jul. 2023, doi: 10.1007/S11831-023-09922-Z.

[11] S. Bird *et al.*, "Natural language processing with Python: analyzing text with the natural language toolkit," 2009, Accessed: May 05, 2024. [Online]. Available: https://books.google.com/books?hl=tr&lr=&id=KGIbfiiP1i4C&oi=fnd&pg=PR5&dq=Steven+Bird,+Ewan+Klein,+and+Edward+Loper+(2009).+Natural+Language+Processing+with+Python.+O'Reilly+Media+Inc.+https://www.nltk.org/book/&ots=Y5zjE4JDJ-&sig=23ju8nwX2UOQcxrfCoy4r4xqUlE.

[12] M. Naeem, F. Rustam, A. Mehmood, … I. A.-P. C., and undefined 2022, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," *peerj.com*, Accessed: May 05, 2024. [Online]. Available: https://peerj.com/articles/cs-914/.

[13] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[14] I. Rish, "An empirical study of the naive Bayes classifier," *cc.gatech.edu*, 2001, Accessed: Aug. 28, 2021. [Online]. Available: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf.

[15] B. Ghojogh and M. Crowley, "Linear and Quadratic Discriminant Analysis: Tutorial," Jun. 2019, Accessed: May 02, 2024. [Online]. Available: http://arxiv.org/abs/1906.02590.

[16] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[17] C. Peng, K. Lee, G. I.-T. journal of educational, and undefined 2002, "An introduction to logistic regression analysis and reporting," *Taylor & Francis*, vol. 96, no. 1, pp. 3–14, 2002, doi: 10.1080/00220670209598786.

[18] M. P.-I. journal of remote sensing and undefined 2005, "Random forest classifier for remote sensing classification," *Taylor & Francis*, vol. 26, no. 1, pp. 217–222, Jan. 2005, doi: 10.1080/01431160412331269698.

[19] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/S10462-020-09896-5.