Factoid vs. Non-factoid Question Identification: An Ensemble Learning Approach

Alaa Mohasseb[1] [a] and Andreas Kanavos[2] [b]
[1]*School of Computing, University of Portsmouth, Portsmouth, U.K.*
[2]*Department of Digital Media and Communication, Ionian University, Kefalonia, Greece*

Keywords: Question Classification, Grammatical Features, Factoid Questions, Information Retrieval, Machine Learning, Ensemble Learning.

Abstract: Question Classification is one of the most important applications of information retrieval. Identifying the correct question type constitutes the main step to enhance the performance of question answering systems. However, distinguishing between factoid and non-factoid questions is considered a challenging problem. In this paper, a grammatical based framework has been adapted for question identification. Ensemble Learning models were used for the classification process in which experimental results show that the combination of question grammatical features along with the ensemble learning models helped in achieving a good level of accuracy.

1 INTRODUCTION

Question Classification is one of the most important applications in the information retrieval area. Questions mis-classification is what mostly affects the performance of question answering systems (Moldovan et al., 2003); to generate the correct answers to the users, it is important to be able to distinguish between the different type of questions.

Distinguishing between factoid and non-factoid questions is considered a very challenging topic. According to (Li et al., 2008), it is difficult to classify ”wh-” questions into semantic categories compared to other types in question answering systems. In addition, to obtain an accurate question, a corresponding classifier feature selection is important (Huang et al., 2008). Different studies classified questions using features such as bag-of-words (Li et al., 2005; Mishra et al., 2013; Yen et al., 2013; Zhan and Shen, 2012), semantic and syntactic features (Hardy and Cheah, 2013; Song et al., 2011; Yen et al., 2013), uni-gram and word shape features (Huang et al., 2008) as well as grammatical and domain-specific grammatical features (Mohasseb et al., 2018b; Mohasseb et al., 2019).

During the last decades, the development of ensemble learning algorithms and techniques has gained a significant attention from both scientific and industrial community (Brown, 2010; Pintelas and Livieris, 2020; Polikar, 2012). The basic intuition behind these methods is the combination of a set of diverse prediction models for obtaining a composite global model, which produces accurate and reliable predictions or estimates. Theoretical and experimental evidence proved that ensemble models provide considerably better prediction performance than single models (Dietterich, 2002). Along this line, a variety of ensemble learning methodologies and techniques have been proposed and implemented their application in various classification and regression problems of the real word (Livieris et al., 2018; Livieris et al., 2019).

In this paper, a grammatical based framework has been employed for question categorization. Ensemble learning models were used for the classification process in which experimental results show that these features combined with these models helped in achieving a good level of accuracy.

The aim of the research presented in this paper is to: *”Evaluate the impact of combining grammatical features and domain-specific grammatical features with ensemble learning algorithms on the classification accuracy and the identification of Factoid and Non-Factoid questions.”*

The rest of the paper is organised as follows. Section 2 outlines the previous work in question classification using different machine learning algorithms.

[a] https://orcid.org/0000-0003-2671-2199
[b] https://orcid.org/0000-0002-9964-4134

Mohasseb, A. and Kanavos, A.
Factoid vs. Non-factoid Question Identification: An Ensemble Learning Approach.
DOI: 10.5220/0011525900003318
In *Proceedings of the 18th International Conference on Web Information Systems and Technologies (WEBIST 2022)*, pages 265-271
ISBN: 978-989-758-613-2; ISSN: 2184-3252

Section 3 describes the approach and the grammatical features used, while the results are discussed in Section 4. Finally, Section 5 concludes the paper and outlines directions for future work.

2 RELATED WORK

In this section, we review previous work on question identification methods. Recent studies proposed question classification approaches by using different machine learning algorithms. In (Golzari et al., 2022), a method was proposed using the feature selection and ensemble classification combined with the Gravitational Search Algorithm. Similarly, a method was introduced in (Van-Tu and Anh-Cuong, 2016) using feature selection algorithm to determine appropriate features corresponding to different question types.

Additionally, authors in (Jiang et al., 2021) used methods such as word segmentation, Part-Of-Speech (POS) Tagging and Named Entity Recognition (NER) for feature extraction. In addition, for question classification, the Support Vector Machine (SVM) and Random Forest algorithms were used. Results showed that SVM and Random Forest methods achieved good results compared to ensemble learning and hierarchical classification methods. In (Li et al., 2005), authors combined statistic and rule classifiers with different classifiers and multiple classifier combination methods. Moreover, many features such as dependency structure, wordnet synsets, bag-of-words, and bi-grams were used with a number of kernel functions. Moreover, in (Metzler and Croft, 2005), a statistical classifier was proposed based on SVMs.

SVMs were also used in (May and Steinberg, 2004; Mishra et al., 2013; Xu et al., 2016); specifically, in (May and Steinberg, 2004), SVM has been used with different classifiers such as MaxEnt, Naive Bayes and Decision Tree for primary and secondary classification. In addition, a question classification method using SVM in addition to k-Nearest Neighbor and Naive Bayes, was employed in (Mishra et al., 2013). The proposed approach also uses features such as bag-of-words, n-grams as well as lexical, syntactic and semantic features. A similar approach in utilized in (Xu et al., 2016), where an SVM-based approach incorporating dependency relations and high-frequency words for question classification, was introduced. Finally, Bidirectional Long-Short Term Memory (Bi-LSTM) were used in (Anhar et al., 2019) for question classification. The classification results showed that Bi-LSTM achieved higher accuracy compared to basic LSTM and Recurrent Neural Network (RNN).

3 ENSEMBLE LEARNING APPROACH

We employ a grammar-based framework for Question Categorization and Classification (GQCC), which was introduced in (Mohasseb et al., 2018b). The framework consists of three phases:

Phase I: Question Analysis. The question is initially analyzed by identifying each of the keywords and phrases in the question to help generate the grammatical rule. After this step, a question domain-specific grammar will be created. This implementation will be done using a simple version of the English grammar combined with domain-specific grammatical categories.

Phase II: Parsing and Mapping. In order to transform each question into its grammatical structure, each question is parsed and tagged using grammatical features combined with domain-related information.

Phase III: Question Classification. In this phase, a model for automatic classification is built and tested.

In this paper, the dataset and grammatical features generated from (Mohasseb et al., 2018b), were employed.

3.1 Dataset

The dataset consists of $1,160$ questions that were randomly selected from the following three different sources:

1. Yahoo Non-Factoid Question Dataset[3]
2. TREC 2007 Question Answering Data[4]
3. Wikipedia dataset[5] (Smith et al., 2008)

Each question in this dataset is classified into six different categories, which are: causal, choice, confirmation (Yes-No questions), factoid (”Wh-” questions), hypothetical and list. These categories are based on the question types in English and the classification is based on types of questions asked by users and the answers given.

For the objective of investigating the impact of the ensemble learning model to distinguish between Factoid and Non-Factoid questions, a new label was created, entitled non-factoid which consists of the five question types, namely causal, choice, confirmation, hypothetical and list. Their distribution is given in Table 1.

[3]https://ciir.cs.umass.edu/downloads/nfL6/
[4]http://trec.nist.gov/data/qa/t2007_qadata.html
[5]https://www.cs.cmu.edu/~ark/QA-data

Table 1: Data Distribution.

Question Type	Number of Questions
Non-Factoid	473
Causal	31
Choice	12
Confirmation	32
Hypothetical	7
List	101
Factoid	687

3.2 Question Grammatical Structure

The main objective of using the question grammatical features is the utilization of the question structure by considering general and domain-specific grammatical categories (Mohasseb et al., 2018b).

One limitation of the aforementioned methodologies, which were introduced so far, is that they use features selection approaches to reduce the number of input variables. As a result, these approaches do not take into account the grammatical structure of the questions. Question characteristics may vary; for example, some questions could be short while other questions might have more than one meaning, which could cause ambiguity, therefore using only a selection of features is not enough. Also, two questions might have exactly the same set of terms but may reflect different intents. Therefore, the classification of the questions using their grammatical structure in addition to domain-specific grammatical categories may help in making the classification process more accurate.

Grammatical features were used to transform the questions by using the grammar into a new representation as a series of grammatical terms. The grammatical features consist of Verb, Noun, Determiner, Adjective, Adverb, Preposition, and Conjunction in addition to question words such as ”How”, ”Who”, ”When”, ”Where”, ”What”, ”Why”, ”Whose” and ”Which”. Furthermore, the grammatical features consist of word classes like Noun and Verbs. More to the point, nouns can have sub-classes, such as Common Nouns, Proper Nouns, Pronouns, and Numeral Nouns; the same stands for verbs, which can have sub-classes, such as Action Verbs, Linking Verbs and Auxiliary Verbs. What is more, the grammatical features consist of other features as well, such as Singular (e.g. Common Noun - Other - Singular) and Plural terms (e.g. Common Noun - Other - Plural). Table 2 provides the list of the grammatical terms and their abbreviation.

Furthermore, domain-specific grammatical features related to question-answering were taken into consideration, which correspond to topics such as Events, Entertainment, History and News, Health Terms, Geographical Areas, Places and Buildings as shown in Table 3 (Mohasseb et al., 2018b).

Table 2: Grammatical Features.

Grammatical Feature	Abbreviation
Verbs	V
Action Verbs	AV
Auxiliary Verb	$AuxV$
Linking Verbs	LV
Adjective	Adj
Adverb	Adv
Determiner	D
Conjunction	$Conj$
Preposition	P
Noun	N
Pronoun	$Pron$
Numeral Numbers	NN
Ordinal Numbers	NN_O
Cardinal Numbers	NN_C
Proper Nouns	PN
Common Noun	CN
Common Noun - Other - Singular	CN_{OS}
Common Noun - Other - Plural	CN_{OP}
Question Words	QW
How	QW_{How}
What	QW_{What}
When	QW_{When}
Where	QW_{Where}
Who	QW_{Who}
Which	QW_{Which}

These grammatical features and structures will be used in the question type identification, since each factoid and non-factoid question type has a certain structure. The different feature representations help in distinguishing between different question types as shown in Table 4.

3.2.1 Question Grammatical Structure Example

The following example *”what are the symptoms of Dementia”* will illustrate how these features are used:

All terms in the questions will be extracted by parsing the following question:

Question: What are the symptoms of Dementia?

The terms extracted are ”What”, ”are”, ”the”, ”symptoms”, ”of”, ”Dementia”.

After the parsing process, each term in the question will be tagged to one of the grammatical features and domain-specific grammatical features, such as:

- *What* = QW_{What}
- *are* = LV
- *the* = D

Table 3: Domain Specific Grammatical Features.

Domain specific Features	Abbreviation
Celebrities Name	PN_C
Entertainment	PN_{Ent}
Newspapers, Magazines, Documents, Books	PN_{BDN}
Events	PN_E
Companies Name	PN_{CO}
Geographical Areas	PN_G
Places and Buildings	PN_{PB}
Institutions, Associations, Clubs, Foundations and Organizations	PN_{IOG}
Brand Names	PN_{BN}
Software and Applications	PN_{SA}
Products	PN_P
History and News	PN_{HN}
Religious Terms	PN_R
Holidays, Days, Months	PN_{HMD}
Health Terms	PN_{HLT}
Science Terms	PN_S
Database and Servers	CN_{DBS}
Advice	CN_A
Entertainment	CN_{Ent}
History and News	CN_{HN}
Site, Website, URL	CN_{SWU}
Health Terms	CN_{HLT}

Table 4: Grammatical Features that Identify Question Types.

Questions	Grammatical Features
Factoid	Question Words such as What, Where, When, Which, Why, Who, How
Non-Factoid	Conjunction (OR), Linking Verbs, Auxiliary Verbs, Plural Common Nouns, Question Words such as What, Which, Who, Why, How

- *symptoms* = CN_{OP}
- *of* = P
- *Dementia* = CN_{HLT}

After tagging each term in the question, the pattern is formulated as illustrated below:

Pattern: $QW_{What} + LV + D + CN_{OP} + P + CN_{HLT}$

The question grammatical feature in each question will be used to identify the question type. As a result, this will produce the final classification of each question. In the given example, the question will be classified as *Non-Factoid*.

3.3 Factoid vs. Non-factoid Question Identification

The algorithms employed and examined in our paper were the following in order to address three particular aspects; for the identification of factoid and non-factoid questions, for the evaluation of using the domain-specific grammatical features with the ensemble learning models and for measuring the accuracy of the classification.

- **Random Forest (RF)** is an ensemble learning method which constructs a multitude of decision trees at the training time in which each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. For classification tasks, the output of the random forest is the class selected by most trees (Breiman, 2001; Ho, 1995).
- **Voting** combines different machine learning classifiers and uses a majority vote or the average predicted probabilities to predict the class labels. Specifically, in this paper, majority vote was utilised. The objective of this method is to improve model performance by using multiple models. In majority vote, the predicted class label is the class label that represents the majority of the class labels predicted by each individual classifier.

  Regarding this classifier, four different models were built and examined in the experiments, namely:

  1. **Voting Model 1 (VM1)** which consists of the following algorithms; Decision Tree, Support Vector Machine and K-Nearest Neighbour (DT, SVM, KNN).
  2. **Voting Model 2 (VM2)** which consists of the following algorithms; Naive Bayes, Decision Tree and K-Nearest Neighbour (NB, DT, KNN).
  3. **Voting Model 3 (VM3)** which consists of the following algorithms; Naive Bayes, Support Vector Machine and K-Nearest Neighbour (NB, SVM, KNN).
  4. **Voting Model 4 (VM4)** which consists of the following algorithms; Naive Bayes, Decision Tree and Support Vector Machine (NB, DT, SVM).
- **Naive Bayes (NB)** is a probabilistic classifier based on applying Bayes' theorem with the assumption that the features occur independently in terms of each other inside a class (Rennie et al., 2003). This classifier has been widely used in text classification because it is fast and easy to implement (Mitchell, 1997).

- **Support Vector Machine (SVM)** uses a hyperplane to separate the data. The objective of this algorithm is to select a hyperplane with the maximum possible margin between support vectors in the given dataset (Cortes and Vapnik, 1995). The implementation of SVM has been very effective in text categorization and predication problems since it can eliminate the need for feature selection making the application of text categorization considerably easier. Furthermore, it does not require any parameter tuning since this classifier can automatically identify good parameter settings (Joachims, 1998).
- **Decision Tree (DT)** is a non-parametric method, which learns simple decision rules inferred from data features to create a model that predicts the value of a target variable.
- **K-Nearest Neighbour (KNN)** also constitutes a non-parametric and instance-based learning algorithm. This algorithm is based on a similarity measure, namely the distance function. KNN forms a majority vote between the k points and then, similarity is defined according to a distance metric between two data points. In the experiments, the value of k was equal to 3.
- **AdaBoost** is a meta-estimator, which fits a sequence of weak learners on repeatedly modified versions of the data. In following, it combines the predictions through a weighted majority vote (or sum) to produce the final prediction (Freund and Schapire, 1997).
- **Bagged DT** is also a meta-estimator that fits each base classifiers on random subsets of the original dataset. This method generates multiple versions of a predictor and uses these in order to produce an aggregated predictor (Breiman, 1996). This method can be as well used to reduce the variance of a decision tree.

4 EXPERIMENTAL STUDY AND RESULTS

In the experimental study, we investigate the ability of the ensemble learning models to distinguish between different question types based on grammatical features. To assess the performance of grammatical features and ensemble learning classifiers, several experiments have been conducted. The experiments were set up using the typical 10-fold cross validation.

Table 5 presents the accuracy of classification performance of the ensemble learning models. Additionally, Table 6 outlines the classification performance details, which are Precision, Recall and F-Measure, of the classifiers that have been examined. The results prove that the use of grammatical features combined with ensemble learning algorithms achieve a high accuracy.

Concretely, Bagged DT achieved the highest accuracy, with value equals to 89% in distinguishing between factoid and non-factoid questions while VM2 has the lowest accuracy, e.g. 79%. Moreover, Table 5 shows that algorithms such as RF, VM1 and Bagged DT are more effective in the identification and classification of factoid questions, whereas VM2 and VM4 classifiers are more accurate in the identification and classification of non-factoid questions.

In addition, regarding Random Forest, this classifier achieved a good performance in classifying factoid questions with value equals to 79%; however it achieved lower recall performance for non-factoid questions with value equals to 78%. Similarly, VM1 and VM3 achieved good results in classifying factoid questions but achieve lower recall values, e.g. 71% and 73% respectively, for non-factoid questions. On the contrary, VM2 and VM4 achieved better results in classifying non-factoid questions, whereas lower recall values, e.g. 74% and 81% respectively, for factoid questions were obtained.

In previous works (Mohasseb et al., 2018a; Mohasseb et al., 2018b), algorithms such as KNN, SVM and NB were combined with grammatical features and domain-specific grammatical features. Specifically, in (Mohasseb et al., 2018a), KNN achieved an accuracy value equals to 83.7%, while in (Mohasseb et al., 2018b), SVM and NB achieved an accuracy of 88.6% and 83.5% respectively. This indicates that the combination of domain-specific grammatical features with ensemble learning algorithms improved the classification accuracy and enabled the machine learning algorithms to better differentiate between factoid questions and non-factoid questions. In addition, nearly all the algorithms achieved an accurate performance and classification accuracy.

The following points summarise the above observations:

- It is clear from results that non-factoid questions was the most difficult question type to predict.
- The classification accuracy and the predication of non-factoid questions was affected by the imbalance of the dataset categories as shown in Table 1.
- Common grammatical features between the two type of questions such as question words what, which, who, why and how affected the identification accuracy of factoid and non-factoid questions

Table 5: Accuracy of the Ensemble Learning Models.

Ensemble Learning Model	*Non – Factoid*	*Factoid*	*avg/Total*
Random Forest	78%	93%	88%
VM 1 (DT, SVM, KNN)	71%	93%	85%
VM 2 (NB, DT, KNN)	87%	74%	79%
VM 3 (NB, SVM, KNN)	73%	84%	80%
VM 4 (NB, DT, SVM)	83%	81%	82%
AdaBoost	80%	90%	86%
Bagged DT	80%	93%	89%

since question words are the main grammatical features that identify factoid questions, as shown in Table 4.

- Bagged DT achieved the highest accuracy, with value equals to 89% in distinguishing between factoid and non-factoid questions.
- Our results showed that algorithms such as RF, VM1 and Bagged DT are more suitable for the identification of factoid questions, whereas VM2 and VM4 classifiers are more suitable for the identification of non-factoid questions.
- Ensemble learning algorithms improved the classification accuracy.
- Domain-specific grammatical features helped in differentiating between factoid questions and non-factoid questions.

Table 6: Classification Performance Details.

Question Type	*Precision*	*Recall*	*F1-Score*
Random Forest			
Non-Factoid	87%	78%	82%
Factoid	88%	93%	90%
VM 1 (DT, SVM, KNN)			
Non-Factoid	85%	71%	78%
Factoid	85%	93%	89%
VM 2 (NB, DT, KNN)			
Non-Factoid	66%	87%	75%
Factoid	91%	74%	82%
VM 3 (NB, SVM, KNN)			
Non-Factoid	73%	73%	73%
Factoid	84%	84%	84%
VM 4 (NB, DT, SVM)			
Non-Factoid	72%	83%	77%
Factoid	89%	81%	85%
AdaBoost			
Non-Factoid	82%	80%	81%
Factoid	89%	90%	89%
Bagged DT			
Non-Factoid	88%	80%	84%
Factoid	89%	93%	91%

5 CONCLUSIONS AND FUTURE WORK

In this paper, ensemble learning models were examined for the predication of factoid and non-factoid questions, as the proposed framework employs grammatical features and domain-specific grammatical features to utilize the structure of the questions. We have employed ensemble learning models as the composite global model produces accurate and reliable predictions or estimates. It is also proven that ensemble models provide considerably better prediction performance than single models.

The proposed grammar-based framework for Question Categorization and Classification consists of three phases, namely, Question Analysis, Parsing and Mapping, as well as Question Classification. Experimental results depict that these features combined with the ensemble learning classifiers helped in achieving a good level of accuracy.

As future work, we aim to investigate the impact on the predication results if imbalance methods were applied on the non-factoid questions combined with different ensemble learning models. Furthermore, another interesting aspect is the use of other component classifiers in the ensemble and enhance our proposed framework with more sophisticated and theoretically sound criteria for the development of an advanced weighted voting strategy. New metrics can also be taken into consideration in order to measure the efficiency of our proposed method, such as Roc Analysis.

REFERENCES

Anhar, R., Adji, T. B., and Setiawan, N. A. (2019). Question classification on question-answer system using bidirectional-lstm. In *5th International Conference on Science and Technology (ICST)*, pages 1–5.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, G. (2010). Ensemble learning. In *Encyclopedia of Machine Learning*, pages 312–320.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dietterich, T. G. (2002). Ensemble learning. *The Handbook of Brain Theory and Neural Networks*, 2(1):110–125.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Golzari, S., Sanei, F., Saybani, M. R., Harifi, A., and Basir, M. (2022). Question classification in question answering system using combination of ensemble classification and feature selection. *Journal of Artificial Intelligence and Data Mining (JAIDM)*, 10(1):15–24.

Hardy, H. and Cheah, Y.-N. (2013). Question classification using extreme learning machine on semantic features. *Journal of ICT Research and Applications*, 7(1):36–58.

Ho, T. K. (1995). Random decision forests. In *3ird International Conference on Document Analysis and Recognition (ICDAR)*, pages 278–282.

Huang, Z., Thint, M., and Qin, Z. (2008). Question classification using head words and their hypernyms. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 927–936.

Jiang, Y., Zhang, X., Jia, W., and Xu, L. (2021). Answer classification via machine learning in community question answering. *Journal of Artificial Intelligence*, 3(4):163–169.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *10th European Conference on Machine Learning (ECML)*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142.

Li, F., Zhang, X., Yuan, J., and Zhu, X. (2008). Classifying what-type questions by head noun tagging. In *22nd International Conference on Computational Linguistics (COLING)*, pages 481–488.

Li, X., Huang, X., and Wu, L. (2005). Question classification using multiple classifiers. In *5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network (ALR/ALRN@IJCNLP)*.

Livieris, I. E., Kanavos, A., Tampakas, V., and Pintelas, P. E. (2019). A weighted voting ensemble self-labeled algorithm for the detection of lung abnormalities from x-rays. *Algorithms*, 12(3):64.

Livieris, I. E., Kiriakidou, N., Kanavos, A., Tampakas, V., and Pintelas, P. E. (2018). On ensemble SSL algorithms for credit scoring problem. *Informatics*, 5(4):40.

May, R. and Steinberg, A. (2004). Building a question classifier for a trec-style question answering system. *The Stanford Natural Language Processing Group, Final Projects*.

Metzler, D. and Croft, W. B. (2005). Analysis of statistical question classification for fact-based questions. *Information Retrieval*, 8(3):481–504.

Mishra, M., Mishra, V. K., and Sharma, H. R. (2013). Question classification using semantic, syntactic and lexical features. *International Journal of Web & Semantic Technology (IJWesT)*, 4(3):39.

Mitchell, T. M. (1997). *Machine Learning, International Edition*. McGraw-Hill Series in Computer Science. McGraw-Hill.

Mohasseb, A., Bader-El-Den, M., and Cocea, M. (2018a). Detecting question intention using a k-nearest neighbor based approach. In *14th International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, volume 520, pages 101–111.

Mohasseb, A., Bader-El-Den, M., and Cocea, M. (2018b). Question categorization and classification using grammar based approach. *Information Processing & Management*, 54(6):1228–1243.

Mohasseb, A., Bader-El-Den, M., and Cocea, M. (2019). Domain specific grammar based classification for factoid questions. In *15th International Conference on Web Information Systems and Technologies (WEBIST)*, pages 177–184.

Moldovan, D. I., Pasca, M., Harabagiu, S. M., and Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS)*, 21(2):133–154.

Pintelas, P. E. and Livieris, I. E. (2020). Special issue on ensemble learning and applications. *Algorithms*, 13(6):140.

Polikar, R. (2012). Ensemble learning. In *Ensemble Machine Learning*, pages 1–34.

Rennie, J. D. M., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *20th International Conference on Machine Learning (ICML))*, pages 616–623.

Smith, N. A., Heilman, M., and Hwa, R. (2008). Question generation as a competitive undergraduate course project. In *NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Song, W., Wenyin, L., Gu, N., Quan, X., and Hao, T. (2011). Automatic categorization of questions for user-interactive question answering. *Information Processing and Management*, 47(2):147–156.

Van-Tu, N. and Anh-Cuong, L. (2016). Improving question classification by feature extraction and selection. *Indian Journal of Science and Technology*, 9(17).

Xu, S., Cheng, G., and Kong, F. (2016). Research on question classification for automatic question answering. In *2016 International Conference on Asian Language Processing (IALP)*, pages 218–221.

Yen, S., Wu, Y., Yang, J., Lee, Y., Lee, C., and Liu, J. (2013). A support vector machine-based context-ranking model for question answering. *Information Sciences*, 224:77–87.

Zhan, W. and Shen, Z. (2012). Syntactic structure feature analysis and classification method research. In *International Conference on Audio, Language and Image Processing (ICALIP)*, pages 1135–1140.