Automatic Text Classification Technology’s Impact on the United Nations

*Kevin Thomas Bradley, Yihan Bao, Praneeth Nooli, Sravanthi Varanasi*

United Nations, Office of Information and Communication Technology, Emerging Technologies[[1]](#footnote-2)

***Abstract***

This paper investigates Automatic Text Classification – the process of assigning topics to textual data. Automatic Text Classification is one of the main techniques in Natural Language Processing problems with broad applications ranging from topic tagging, name entity detection to spam detection. Automatic text classification involves machine learning with the usage of natural language processing to automatically classify the text in a more cost-efficient way. This paper gives an overview of the technology, how it is being adopted generally and how the United Nations is using and could further use this to transform the way it classified text by a human annotator.

# Background

Automatic Text Classification is a system implemented to categorize text in line with a set of topics/categories after a training stage. The target of text classification is to automatically classify the articles or documents into one or more defined topics. It is now widely applied in people’s daily life. Some real-world examples of text classification are Email spam detection, social media trends analysis, customer query automatic tagging, webpage topic classification, news article classification. Automatic text classification can be roughly grouped into three different types: The rule-based system, machine learning based system, and hybrid systems.

# 1.1 History

Many researchers have actively investigated text classification since the 1980s. From the perspective of text analytics, text classification can be treated as preprocessing technology to filter out irrelevant articles from many documents. The task of text classification is to categorize documents into different predefined topics. Due to the huge increase in online textual information, e.g., email contents, web news, as well as a great number of resources for scientific online abstracts, there is a consistently developing demand for text classification. It is an interesting research field about how to accomplish high performance in the task of assigning multiple topics to certain articles in a targeted domain. The task to categorize each article into multiple topics is called multi-label text classification.

The origin of text classification returned back to the early '60s. And 30 years later, the machine learning techniques were successfully applied into the text classification field [1]. Support vector machines (SVM) were successfully applied to text classification technique in 1998. Later, AdaBoost was promoted to handle the multi-label text classification issue in 2000 by Schapire and Singer. And recently, text with multiple topics has been investigated. And at this point, auto text classification is inspired by information retrieval to rank the topic candidates. However, in a text classification problem, a set of definite topics for each article was required not the rankings of topic candidates. Entered 2000, other techniques were raised to solve the auto text classification problem, such as maximum entropy models.

# 1.2 Approaches of Text Classification

***Rule-based method:*** The rule-based method is used when people set the rules themselves and let the machine to implement. Each rule is instructed by a pattern. And these rules will direct the system to categorize the textual data into different groups [2].

For Example, let us consider some news articles which contains topics like Politics, Sports, Business, etc. We clarify the articles into different categories based on “rules”, such as searching keywords in the news article. If “basketball”, “volleyball” and “NBA (keywords in Sports category) appear most compared with other categories' keywords, tag the article with “Sports”.

***Machine learning method:*** Compared with the rule-based method’s setting up rules manually, this method uses word vectorization method and machine-learning algorithms to classify the articles or documents. The machine learning method is done based on past observations. By using pre-labeled examples as the training dataset, an algorithm can be trained with proper parameters to get the relationship between textual data and the output labels.

For instance, a single label text classification problem will go through the same procedure as the normal machine learning project: data preparation (including word vectorization), feature engineering, training model evaluation and selection and prediction part.

***Hybrid method:*** Hybrids methods combine the machine learning techniques and rule-based method together. The hybrid system will adjust the results of machine learning method by rule-based method, which can further improve the results or avoid the specific conflicting tags that can’t be correctly modeled by the machine learning classifier [3].

For example, to correctly tagging the labels, the hybrid method first applies a machine learning algorithm and get the predicted labels as the basic labels for each document, and then for correcting the basic labels that are mis-predicted at the first step, rule-based method is implemented to revise the output label at the first step.

# Implementations

Businesses these days are driving towards text classification for structuring the text to enhance decision making and automate the process of sentiment analysis, labeling a topic, detecting spam and intent. Text classification can be implemented over unstructured data which exists in the form of text exists in emails, chats, web pages, social media, etc. Generally, text classification problems are categorized into single-label or multi-label problems.

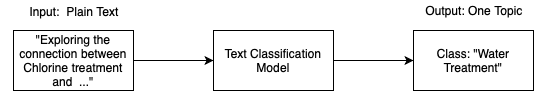
* 1. Single-label text classification

Single-label text classification assumes that the predefined data categories are mutually exclusive, and each data point can belong to exactly one category. Single-label text classification is categorized into two categories:

***Binary text classification:*** It is a task of classifying data points of a given set into one or two groups.

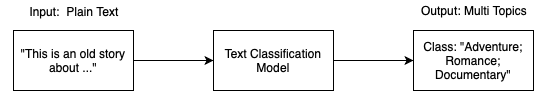
***Multi-class text classification***:  It is a task of classifying data points of a given set into three or more groups.

Binary classification is the simplest case of the single-label problem. To date, many classification methods, such as Naïve Bayes, SVM, and Logistic Regression, have been developed to address the single-label text classification problem and here is a small example of it.



2.2 Multi-label text classification

Multi-label classification originated from the investigation of text categorization problem, where each document may belong to several predefined topics simultaneously with multi-label text classification, the data categories may not be either mutually exclusive or conditionally independent and each data point can belong to multiple categories simultaneously. For example, a certain type of fiction can belong to both “adventure novel” and “narrative novel” at the same time. Multi-label text classification is very common in the area if document analysis and information retrieval. And here is an example of multi-label text classification.



2.3 Text Classification Algorithms

The text classification problem is one type of classification problem in machine learning, so the algorithms that can be used on classification can also be used for text classification. Here are some of the popular machine learning algorithms for generating automatic text classification models. [4]

***Naive Bayes:*** Naïve Bayes is one implementation for statistical algorithm based on Bayes’ Theorem which is used to perform the computational probability form two events based on the probability of each individual event occurrence. This algorithm is majorly used for spam detection. To understand the Naïve Bayes algorithm let us consider an article must be classified into certain groups, in that case the vectors that represent the article should have the information about the probabilities of each word that appears within the text of the article. This helps the algorithm to compute the similarities between the texts that belong to the same category.

***Support Vector Machines:*** Support Vector Machines: Support Vector machine is another implementation for text classification. This method uses hyperplane to classify the data and provide the best results. Unlike Bayes, SVM does not need much training data to give accurate results. But the computational power needed by SVM is too high to achieve more results. The hyperplane divides the space into two subspaces where one subspace contains the vectors which don’t belong to another group. The groups are the tags that are provided for each of the vectors created.

***Deep Learning:*** Deep Learning: Deep Learning is another implementation for text classification. The deep learning algorithms are classified into Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The Convolutional Neural Networks (CNN) is developed to store the spatial of a problem and to recognize an object. The Recurrent Neural Networks (RNN) are used to solve the difficult machine learning problems that involve a sequence of inputs. This network has connection loops which allow in adding feedback and memory to network over time that help in learning and generalize the sequence of input rather than an individual pattern. The general machine learning algorithms have a threshold while adding more training data and doesn’t give an accurate output. This phenomena of improving the accuracy can be achieved by using the deep learning classifiers as they get better on adding more data.

1. Methods

There have been lots of different approaches to applying text classification, such as Microsoft Azure Machine Learning, Amazon Web Service Machine Learning, and Python programming language. All methods are usually divided into two groups, one is to use a programming language to solve the problem and another is to use platforms and services provided by a different company.

3.1 Python Programming Language

Python is an extremely scalable language with wide data science libraries and natural language toolkit (NLTK) is the most popular library for natural language processing (NLP), which was written in Python and has a big community behind it. And python language is open source and contain lots of build in packages for machine learning techniques.

3. 2 R Programming Language

R is a programming language, which is good at statistical computing. R language is widely used in statistical analysis and data mining. To better analyze the relationship between the data, R will be the best choice.

3.3 Machine Learning Platform

There are many kinds of machine learning platform provided by a different company. For example, Microsoft creates a Machine Learning Studio that can accomplish text classification. The Microsoft Azure Machine Learning Studio [5] is a collaborative, drag-and-drop tool that you can use to build, test, and deploy predictive analytics solutions on your data. And Amazon Web Service [6] also create one section called Machine Learning, which provides many powerful API related to text mining.

1. Practical Uses

Text classification is widely used and applied in today’s technology, especially on the business field to improve the efficiency and help on decision making by predictions it provides.

* 1. Tag contents and products

Text classification can be used for tagging contents by categorizing the content on the webpage so that it will be helpful for the user to have a better experience. Google uses content tagging to crawl the websites and ultimately provide the best results for the user. By applying text classification over the websites and other platforms such as directories, blogs, news agencies to automate and speed up the content extraction and provide accurate results in no time [7].

* 1. Automate Customer Relationship Management

Text classification can also be used to build good customer experience for a growing company so that they can provide excellent customer service with the reviews provided by the customer. With input from the customers, text classification can help build an automated task.

For Example, think of a company that provides customer support in several languages, so the tickets have to be assigned based on the language. To do this a person has to manually identify the ticket and assign it to the respective language representative. So, providing responses to such cases would require a lot of time and manual effort. So, by using text classification we can automate ticket routing and assigning.

* 1. Emergency response system

Text classification can be used for classifying a panic conversation on a social media platform and authorities can give a quick response at that situation. We can monitor the incoming conversations via any social media platform and sense if any critical emergency situations like fire, crime we can send a call to the respective department so that they can handle the situation. This helps in facilitating people with quick communication and respond at the earliest to the emergency which would benefit them.

1. Nuances

Automatic text classification technology specific nuances such as its operation mechanism – classification, domain knowledge such as natural language understanding and its input and output data – text and label (topic).

5.1. Classification

Automatic text classification technique belongs to supervised machine learning techniques – classification. The only difference in text classification and machine learning classification is the input data type. For the traditional machine learning classification issue, the data is numeric but here the text classification problem targets to solve the problem with textual data. They share the same part: classification, which is to group all the input data into several pre-defined labels or here for text classification is called “topics” [8,9].

5.2 Natural Language Understanding

Natural Language Understanding is an important subfield of natural language processing in Artificial Intelligence where we are trying to make computers interact like humans. Trying to make systems understand the way humans can analyze a conversation and respond back to the situation [10].

5.3 Text and Labels

In the automatic text classification problem, input data is textual data and output is called labels meaning the topic for each text record. In single label text classification, the number of labels is one and in multi-label text classification, the number of labels is at least one. Text and labels are both necessary to fulfill the automatic text classification.

1. United Nations Specific Use Cases

The United Nations contains lots of documents that need to be classified according to the pre-defined labels. There are many offices that met the issue with spending on so much time on manually classifying the documents with different topics. For example, the Office of Disaster Risk Reduction and Office for the Coordination of Humanitarian Affairs.

* 1. UNISDR Web Prevention Tools - Articles Classification

Here for the Office of Disaster Risk Reduction, they are in a demand to implement a multi-label text classification of the articles on the websites. Each UNISDR article is related with either one or more disasters from 17 unique hazards that happen around the world, such as avalanche, flood, and wildfire or one or more 34 different themes and issues, for instance, Community-based DRR, Disaster Risk Management, Food Security & Agriculture, etc. There is no explicit specification of the category or topic the document focused on, therefore, to label these articles into categories, currently, UNISDR had been manually reading and tagging them individually.

To save manpower and increase the efficiency of classifying articles, a solution by applying Machine Learning and Natural Language Processing knowledge on this project is highly recommended.

6.2 OCHA Text Classification Service - Reports Classification

For the Office for Coordination Humanitarian Affairs, there are lots of WHS reports needed to be categorized. Each WHS report belongs to one of the 5 core responsibilities and 24 transformation techniques and each transformation is multi-labeled. For example, one report which belongs to E responsibility and 2E transformation, which can be categorized into two labels: UNSC and Other-2A. There is no explicit specification of the category or topic the document focused on, therefore, to label these articles into categories, currently, OCHA had been manually reading and tagging them individually.

By applying the automatic text classification, the complicated reports classification problem can be solved by artificial intelligence techniques. This will be a great help for the OCHA and save a lot of time and manpower on this report categorization.

1. Benefits & Drawbacks

To better apply the automatic text classification technology, it is better to know much details about text classification’s advantages and disadvantages in the field of natural language processing.

* 1. Key Advantages

There are many advantages that automatic text classification has compared with the human. Automatic text classification is in the field of natural language understanding in artificial intelligence, which is using machine learning technology and natural language processing technique to find the pattern inside the textual data and get the correlation between features and labels. It is much powerful when some feature is difficult to detect by mankind.

Considering the time performance and error rate, automatic text classification is much better at these two aspects. In general, the work of tagging and categorization is time-consuming, and mankind will take much time working on the classifying the documents and reports. And when all the documents’ categories are mainly depending on mankind’s personal selection, which doesn’t have criteria and a large amount of work will cause the error tagging rate increase.

* 1. Disadvantages

The disadvantages of automatic text classification are that the target labels should only be pre-defined. Any other labels that are not included in these labels cannot be applied in automatic text classification or additional training need to be done with the new labels.

Another disadvantage of automatic text classification is that if the textual data is less, the models trained by the automatic text classification will reach for a low accuracy, which won’t have a valuable usage to solve the issues.

* 1. Technical Challenges

There are mainly three challenges involved in the automatic text classification problem. First is the imbalanced data problem, for single label text classification, the imbalanced data can be solved by applying SMOTE techniques but for the multi-label text classification’s imbalanced data problem, it is still in a research state. Though there are several methods to deal with the imbalanced data, there isn’t the best solution for it in general. Second is how to calculate the accuracy, we know for the multi-label text classification problem, the textual data will have one or more labels for different documents, so the number of labels does not stick to the same. And the consideration of confusion metrics will be a lot valuable. For example, if the projects care most of False Positive, which means that the label should be 0 but the model predicts it as 1, one should adjust the accuracy algorithm to avoid this situation. And the third technical challenge is that if it needs to give back the response at the real time, how many features are selected and which methods to use for feature selection is very important. Since the more features are selected, the more operation time it needs.

* 1. Associated Costs

For the use of the automatic text classification, the methods of using Python scripts are free and for using the AWS and Microsoft platform, it needs the subscriptions based on different services.

1. Best Practices

Gmail Spam Filter is one of the best practices for single label text classification. It works by using algorithms to detect which words and phrases are most often used in the spam emails. And serval hundred rules are applied to each email that passes Google’ data center. Each rule describes some attributes of spam and has some numerical value associated with it, based on the likelihood that the attribute is spam. The resulting value is the spam score for the message. This score is then tested against a sensitivity threshold set by an individual’s spam filter. And thus, it is categorized as spam or valid email.

* 1. Consideration

The following list is some considerations whilst planning for or undertaking a text classification-based project:

* Full understanding of the problem requirements and determination on whether it is feasible or not to use automatic text classification technique to solve the issue.
* Comprehensive understanding of the various implementations of automatic text classification. One must have in-depth knowledge of the natural language processing technology and machine learning knowledge to apply this technique successfully.
* Usually, the automatic text classification problem will encounter the imbalanced data issue, whether to use maximum entropy method or independent binary classifiers or other methods to solve this issue need to be taken into considerations.
* The appropriate algorithm to evaluate the models’ performance – for the multi-label text classification problem, it is necessary to create an algorithm to calculate the accuracy based on different cases.
* There are three approaches in automatic text classification technique, rules-based, machine-learning based and hybrid methods. All these methods have their own advantages and disadvantages. Select the best method and suits your project most.
  1. Workflow

An automatic text classification system will accept the input data or files and give the outputs label at one time. To create a workflow of automatic text classification, two main parts needed to be completed, one is the training part, and another is the testing part. The training part is very important which will find the proper patterns between input textual data and output labels. And during the testing part, you put the new datasets into the models and get back of labels, which is your results.

1. Conclusion

In conclusion, automated text classification can, will and should help revolutionize the way the United Nations delivers its mandates. Provided in *Section 6 United Nations Specific Use Cases* are some of the many identified uses to provide a flavor of the variability of this technology to ultimately transform the way how people classifying the textual data.

9.1 Current State

In recent years, there has been an exponential growth in the number of complex documents and texts that require a deeper understanding of machine learning methods to be able to accurately classify texts in many applications. Many machine learning approaches have achieved surpassing results in natural language processing. The success of these learning algorithms relies on their capacity to understand complex models and non-linear relationships within data. However, finding suitable structures, architectures, and techniques for text classification is a challenge for researchers. And different text feature extractions, dimensionality reduction methods and evaluations methods need to be selected based on the real-world problem.

9.2 Recommendations

Adoption of the technology can return intangible value to the organization and solve time-consuming issue within seconds.

Due to the time and manpower it saved and the stability of its accuracy, automatic text classification is right now an emerging technology that used to improve efficiency in the existing working processes. Consider initially people manually classifying all the documents and articles by themselves, using automatic text classification technique can improve the efficiency a lot.

This paper does not cover the entire landscape of automatic text classification but has highlighted the most relevant aspects for consideration within the United Nations. Further exploration of the field to assist adopters is necessary, but hopefully, the contents included illuminate the fundamental concepts.

References

[1] <https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/tutorial-TC.html>

[2] <https://www.meaningcloud.com/developer/resources/doc/models>

[3] <https://arxiv.org/abs/1009.4574>

[4] <https://monkeylearn.com/text-classification/>

[5] <https://studio.azureml.net>

[6] <https://aws.amazon.com>

[7] <https://towardsdatascience.com/text-classification-applications-and-use-cases-beab4bfe2e62>

[8] <http://www.nltk.org/book/ch06.html>

[9] [https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/](https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-%20%20%20%20%20%20%20%20%20%20%20%20%20%20text-classification-in-python/)

[10] <https://towardsdatascience.com/an-easy-introduction-to-natural-language-processing-b1e2801291c1>

1. Statements in this paper represent the opinions of the authors and not necessarily the position of the United Nations nor its Office of Information and Communication Technology, Emerging Technologies. [↑](#footnote-ref-2)