sustainability

Article

Constructing Differentiated Educational Materials Using Semantic Annotation for Sustainable Education in IoT Environments

Yongsung Kim, Jihoon Moon and Eenjun Hwang *

School of Electrical Engineering, Korea University, Seoul 02841, Korea; kys1001@korea.ac.kr (Y.K.); johnny89@korea.ac.kr (J.M.)

* Correspondence: ehwang04@korea.ac.kr; Tel.: +82-2-3290-3256

Received: 21 March 2018; Accepted: 16 April 2018; Published: 23 April 2018

Abstract: Recently, Internet of Things (IoT) technology has become a hot trend and is used in a wide variety of fields. For instance, in education, this technology contributes to improving learning efficiency in the class by enabling learners to interact with physical devices and providing appropriate learning content based on this interaction. Such interaction data can be collected through the physical devices to define personal data. In the meanwhile, multimedia contents in this environment usually have a wide variety of formats and standards, making it difficult for computers to understand their meaning and reuse them. This could be a serious obstacle to the effective use or sustainable management of educational contents in IoT-based educational systems. In order to solve this problem, in this paper, we propose a semantic annotation scheme for sustainable computing in the IoT environment. More specifically, we first show how to collect appropriate multimedia contents and interaction data. Next, we calculate the readability of learning materials and define the user readability level to provide appropriate contents to the learners. Finally, we describe our semantic annotation scheme and show how to annotate collected data using our scheme. We implement a prototype system and show that our scheme can achieve efficient management of various learning materials in the IoT-based educational system.

Keywords: internet of things; sustainable computing; semantic annotation; multimedia in education

1. Introduction

In recent years, Internet of Things (IoT) technology has been utilized in a wide variety of fields, and many studies have been done to improve its performance. The concept of IoT was originally proposed by Kevin Ashton in 1999 [1,2]. He referred to IoT as uniquely identifiable interoperable connected objects with radio-frequency identification (RFID) technology [1,3,4]. In another definition, the idea of the IoT is to integrate all devices (electronic devices, mobile devices, vehicles, appliances, and so on) into the network, which can be managed from the web and in turn, provide information in real time and also allow the interaction with people who use it [5].

Such IoT technology is now widely used in various fields such as health care, transportation, electricity, education, and so on [6–9]. In particular, IoT technology used in the field of education aims to enhance the learning efficiency of learners. To achieve such a goal more effectively, various IoT physical devices such as Near Field Communication (NFC), QRcode, Raspberry Pi, Arduino, and mobile devices are used in the class and diverse interaction data with such devices are collected and analyzed [5]. Such interaction data can be used for diverse purposes such as learning path design and learning content recommendations [4,5,10]. So far, many studies have been done to utilize IoT technology in education, which includes IoT-based interaction systems [5], learning content recommendation systems [4], and learning frameworks [11]. In these IoT-based educational systems, various multimedia content available on the web or the social network service (SNS) is provided to the learners through the physical devices and usually they perform diverse actions on the contents and the devices. Such user interaction data can be collected through the devices and analyzed for diverse purposes such as learning analysis and learning content recommendation. One critical problem to this is that various data used in this environment have different formats and standards. This becomes an obstacle for the machine to understand the meaning of the data, which, in turn, makes it difficult for instructors and learners to reuse the data. In other words, sustainable management of such data has not been easy in IoT-based educational systems.

To solve this problem, in this paper, we utilize the semantic web technology to the data generated through the physical devices and the multimedia contents on the web/Social Network Service (SNS). The semantic web technology enables machines to interpret, combine, and use data on the web [12]. In particular, we use the semantic annotation technique to annotate metadata about resources and generate machine-readable descriptions [12]. As a result, the user can easily search, analyze, aggregate, and reuse various types of data [13]. Specifically, we first introduce a few examples to show how this scheme can be used for N-STEAM (science, technology, engineering, arts, and mathematics) education. Second, we group a large volume of Twitter news quickly by using various machine learning methods. Third, science/technology Twitter news contents are further classified using a topic modeling method and the topic of each news item is extracted. Fourth, we calculate the readability of each news content to determine its grade level. Fifth, we create an optimal news data set for N-STEAM education by removing similar or redundant news. Finally, to enrich the semantics of Twitter news, we produce Twitter news in Resource Description Framework (RDF)/Extensible Markup Language (XML) format by semantic annotation using various previously extracted data.

The main contributions of this paper can be summarized as follows:

1. We propose an N-STEAM educational method that combines NIE and STEAM education to utilize SNS news and various multimedia content on the web as learning materials for effective STEAM education.
2. We show how to collect learners' data through various IoT devices and evaluate their levels for differentiated learning. We also show how to evaluate the difficulty of news using a well-known readability formula without a bias against specific metrics.
3. We show how to provide or recommend appropriate learning materials to the learners depending on their levels and navigate them easily.
4. We show how to apply semantic annotation to the collected data so that various types of data in the IoT-based educational system and SNS data on the web can be reused.

The rest of this paper is organized as follows. In Section 2, we discuss the background of this study and the related works. In Section 3, we describe our scheme in detail. In Section 4, we present our experimental results. In Section 5, we briefly discuss the conclusion and outline our future work.

2. Background and Related Works

2.1. Internet of Things and Semantic Web

2.1.1. Semantic Web of Things

So far, various definitions of IoT have been proposed, but there is no universal definition yet for IoT [3,4,14]. IoT technology utilizes various IoT devices from which a very large amount of data is generated. The semantic web is a technology that helps users manage these data more effectively and continuously. It also makes it easy for a computer to understand the data so that the computer can easily process data in a wide variety of formats and standards. Therefore, existing IoT technology needs to be adapted to Semantic Web of Things (SWoT) based on the following three phases [15]. The first phase was to interconnect everything to the internet (Internet of Things). The second phase was to connect things to the web using the standard solutions already adopted in the web (Web of Things). The Web of Things (WoT) allows different things and systems to interact with each other, thereby allowing the composition of more complex services and solutions. The challenge after the Web of Things is to build an SWoT in order to ensure a common understanding. SWoT promises a seamless extension to the IoT allowing integration of both the physical and digital worlds. SWoT is focused on providing a wide scale of interoperability that allows sharing and reuse of these things. Therefore, for sustainable computing in the IoT environments, IoT/WoT needs to be moved to the SWoT.

2.1.2. Semantic Annotation of SNS Data

Semantic annotation transforms unstructured data into a structured representation, which enables application programs to better search, analyze, and aggregate information [13]. Thus far, many studies have been conducted to make semantic annotations of SNS data. For instance, Lösch and Müller [16] devised a method for annotating microblog posts (Twitter) that contain hashtags with related encyclopedia entities. For the annotation, they collected microblog posts containing query hashtags, analyzed the original news content, extracted the topic, and annotated it using a DBpedia entity. Such annotation could be an effective tool for users to quickly grasp the meaning of a hashtag and find a starting point for further exploration of the hashtag context. In this work [16], the authors just used Twitter contents, dates, and Uniform Resource Locators (URLs) to convert tweets to RDF and did not consider any additional information such as the number of retweets or favorites. Abel et al. [17] proposed a method for linking Twitter posts with related news articles to contextualize Twitter activities. They used OpenCalais to extract entities and topics from tweets and news, and the extracted information is used for semantic enrichment by annotating tweets and news. The connection between the semantically enriched news articles and the Twitter posts makes it possible to construct a rich RDF graph. However, the authors did not present a concrete way to produce such RDF, but focused on simply linking the Twitter posts and related news. In [18], the authors proposed a solution for determining what a microblog post is about through semantic linking: they added semantics to a post by automatically identifying concepts that are semantically related to it and generating links to the corresponding Wikipedia articles. To do this, they used a machine learning technique with a set of innovative features and showed that significant accuracy improvement could be achieved in the concept extraction.

2.2. Pedagogical Theory

2.2.1. STEAM Education

Recently, STEAM (science, technology, engineering, arts, and mathematics) education has been spreading and used in various regions and grades [19–21]. STEAM is an emerging educational model of how the traditional academic subjects of science, technology, engineering, arts, and mathematics can be organized into a framework by which to plan an integrative curriculum [22]. STEAM is based on STEM (science, technology, engineering, and mathematics) education, which can be defined in two ways: In the first definition, STEM emphasizes the individual fields of science, technology, engineering, and mathematics subjects rather than the integration of four subjects. In the second definition, STEM is a new teaching method that emphasizes the integration of four subjects rather than the characteristics of each subject and that aims to improve the efficiency of teaching and learning by intentionally integrating those subjects. STEAM education is similar to the latter definition and emphasizes the integration of those five subjects. STEAM education does not entail a part of education but refers to the overall paradigm from professional learning to lifelong learning [19,22]. There are several interesting approaches for STEAM education. One noteworthy approach is to utilize the latest Information Technology (IT) in STEAM education. In particular, this could be an attractive educational method to digital-generation students since they are familiar with the latest contents of science and technology and have interest in catching up with the trends. However, existing STEAM education has been very limited in reflecting fast changes in science, technology, and engineering. As a result, the students might lose interest or get bored in the class [19]. To solve these problems, we proposed a semantic enrichment scheme for the latest science/technology Twitter news for differentiated STEAM education lessons. Our proposed scheme enables the learners to browse news of desired topics and their relevant materials, even filtered by the user level [23].

2.2.2. NIE (News/Newspaper in Education)

NIE is an educational method that utilizes newspapers as an educational resource in the classroom [24]. It has been shown that NIE can improve the learning abilities of students including critical-thinking ability, literacy, organizational skill, and language skill, and helps to increase the knowledge on current events [25–27]. Oliveras et al. [27] confirmed that the critical thinking ability of the learner was improved when reading science newspapers through various critical thinking activities in science class. Wang [28] found that the use of newspapers in science education has proven to promote the science learning performance of students including their learning attitudes, interest in science classes, and science reading attitudes. Because of its many advantages in class, NIE is a good educational method for improving the learners' diverse abilities.

With the rapid development of IT technology, various online news services have become available and replaced traditional newspapers [25]. Accordingly, NIE is shifting to online news instead of traditional newspapers. This is called digital NIE, and the term NIE is more commonly used to refer to news in education rather than to newspaper in education [29]. Choo [26] proposed the first online NIE model based on online news. In this model, the authors focused on using online news with multimedia content and interactive features. In addition, they presented various teaching methods that could be used in NIE. However, they did not consider the popularity of the news nor did they mention how to select the appropriate news from among the news sources for learners. In our previous studies [30,31], we proposed a platform called TNIE (Twitter News in Education) that utilizes Twitter in NIE. In this study, Twitter news was classified according to its topic and classified news topics were provided to learners through various visualization tools. In addition, learners were able to conduct various discussions and cooperative learning through classified news. However, since the study did not consider the readability of Twitter news, news topics that were provided to the learners could be either too easy or too difficult. In addition, TNIE could not provide news-related materials. In our previous study [4], we proposed an IoT-based scheme for recommending learner-customized multimedia contents for supporting NIE lessons. To provide learners with various types of multimedia data (news, video, and LD) on the web, we utilized data from devices such as Raspberry Pi and mobile devices. We have also published these heterogeneous multimedia data in nonsemantic and semantic data formats, allowing users to utilize them regardless of the data format.

2.3. Tweet Classification

Twitter is a popular online news and social network service where users can post and interact with messages of up to 140 characters (termed tweets) to each other [32]. A large amount of news is generated every second on Twitter and disseminated quickly to the public [30]. However, owing to the size limitation of Twitter news, it has difficulty in delivering detailed information.

Sriram et al. [33] proposed a method for efficiently classifying tweets by using a minimal set of features to represent the short text. They used a small set of features, namely, the 8F feature set to classify incoming tweets into five generic categories: news, opinions, deals, events, and private messages. Rosa et al. [34] showed how to automatically cluster and classify tweets into different categories using unsupervised and supervised methods. They found that unsupervised methods tend to classify tweets according to language similarity rather than to topical coherence. However, standard supervised methods actually work well on these short and noisy data. Sankaranarayanan et al. [32] suggested the use of Twitter to build a news processing system called TwitterStand. The goal of this work was to demonstrate how to use Twitter to automatically obtain breaking news from the tweets posted by Twitter users and to provide a map interface for reading this news since the geographic location of the user as well as the geographic terms comprising the tweets play an important role in clustering tweets and in establishing the clusters' geographic foci.

2.4. Readability Analysis

Readability indicates the degree of ease with which a person can read and understand written materials. It can be measured by a mathematical formula that calculates a grade level according to the length of words, length of sentences, complexity of word use, etc. [35]. Generally, easy-reading text improves comprehension, retention, and reading speed, and the average reading level of the US adult population is at the eighth-grade level [36,37]. Many readability measurement formulas have been developed, and various studies have been conducted to measure the readability of documents in various areas [23].

Freda [35] measured the readability levels of the American Academy of Pediatrics (AAP)'s patient education brochures using the Flesch–Kincaid and Simple Measure of Gobbledygook (SMOG) formula. The result showed that some brochures have acceptably low levels of readability, but at least half were written at higher than acceptable readability levels for the general public. This study also demonstrated statistically significant variability between the two different readability formulas. Ghose and Ipeirotis [36] extracted features such as readability and spelling errors in the product review text and identified how they affected product sales. In this study, the readability level of the review text was calculated using the Automated Readability Index (ARI), Coleman–Liau Index, Flesch Reading Ease, Flesch–Kincaid Grade Level, Gunning Fog Index, and SMOG. They found that in some product reviews, such as of digital cameras, higher readability scores are associated with higher sales. They also found that an increase in the readability of reviews has a positive and statistical impact on review helpfulness.

3. Proposed Scheme

In this section, we first present a scenario of an N-STEAM education lesson, and then describe the overall architecture of our proposed system and some of the implementation details.

3.1. Scenario of an N-STEAM Education Lesson

- Project Name: Design and production of a car model that our family will ride in 2025 using a foam board!
- Conditions:
  - Reflect the latest science/technology (based on current time).
  - Apply one or more new technologies that will emerge in the future based on the latest science/technology.
  - Produce a car body that has less wind resistance.
  - Produce a car that gives priority to family safety.
  - Perform design and production within 5 h in total.
  - Include all the elements of STEAM in the final product.
- Project progress order:
  - Problem Identification: Each team confirms the above problem situation and conditions given by the instructor.
  - Solution Search: **Each team gathers news on the latest automotive technology through Twitter and collects various relevant learning materials through the web.** In other words, learners collect data on the latest science/technology topics that have recently attracted much attention from people such as electric cars and autonomous vehicles. They will use

these collected materials to conceive those components used for the cars that will emerge in 2025 (related subjects: S, T, E, and A elements).

- Design: Each team designs a car creatively, reflecting the proposed conditions (related subjects: S, T, E, A, and M elements).
- Production: Each team produces a car model according to the design using a foam board. Learners can improve teamwork through collaborative learning (related subjects: S, T, E, and A elements).
- Evaluation: The instructor evaluates the car model produced by each team by considering how many latest science/technology elements are used. The finished products are also evaluated by peers.

- How to use learning materials:

    - The learner is provided with the Raspberry Pi and camera module to scan the QR code presented by the instructor. Then, a news article is provided to the learner to check the level of the learner. The learner reads it, checks the degree of difficulty, and sends it to the server. At this point, the instructor is notified of the learner's level. In the class, learners can formulate queries to search for news of desired topics and related multimedia contents. Since all of the learning materials are semantically annotated, searching for learning materials and navigating through them can be done very efficiently.

3.2. Overall System Architecture

Our system consists of three main components: DC (Data Collector), TNA (Twitter News Analyzer), and OC (Ontology Component). DC selects the most popular "news Twitter accounts" and collects Twitter news. It also collects data for educational video material and physical devices. TNA classifies science/technology news through various machine learning methods and extracts various topics through a topic modeling method. In addition, it calculates the readability level of all the classified news to provide news suitable for the learner's level. Finally, OC enables the instructor and learner to quickly and easily search the news and related materials needed for N-STEAM education according to the semantic annotation on Twitter news and related materials. Figure 1 visualizes the flow of our proposed method.

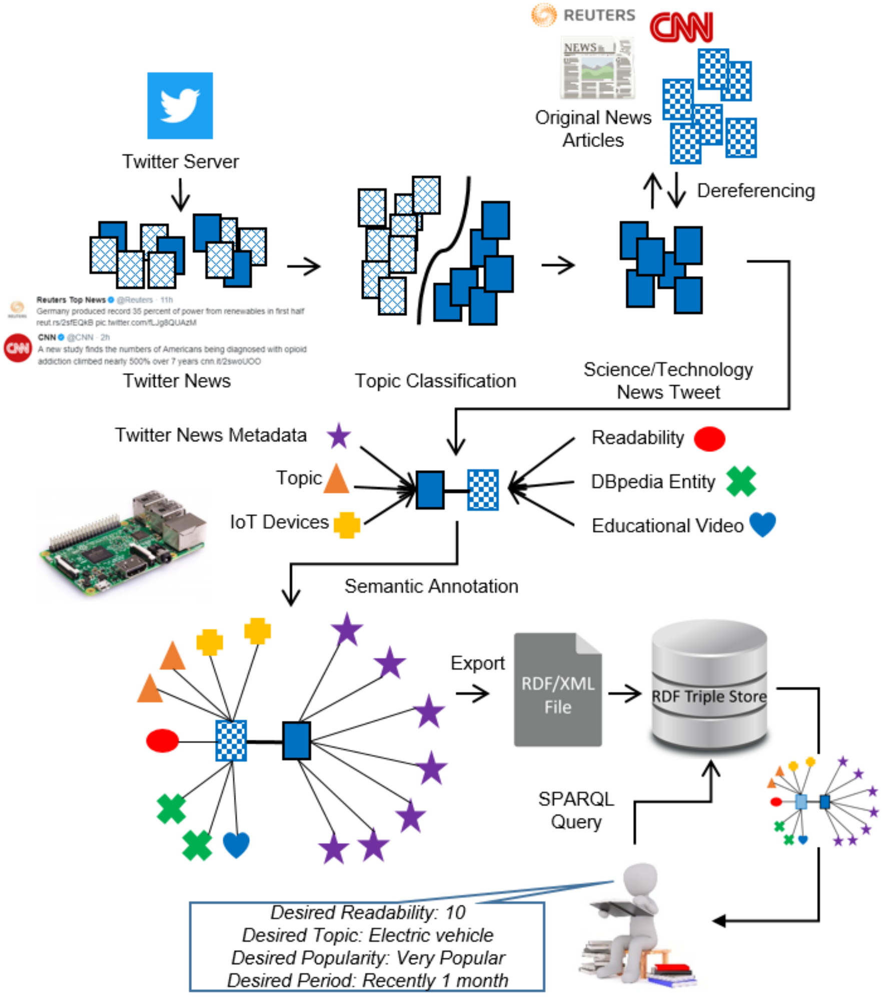
3.3. DC (Data Collector)

3.3.1. Twitter News Collection and Preprocessing

A huge amount of news articles are generated every day around the world, and these news articles are spread through various channels. Most people use computers or smart phones to read news articles. Most newspaper companies run an online news website and upload articles on a wide variety of topics. However, since there is no indicator of public interest in a particular news item on a typical news website, it is very hard to figure out how much interest the public has in specific news. Meanwhile, most of these companies also post breaking news and hot topic news through their Twitter account, which plays an important role in attracting the attention of the public by spreading the news quickly. Utilizing the various features of a Twitter message such as retweet and favorite allows the important properties of the news such as its popularity to be figured out easily.

Considering the large volume and small size of tweets, filtering news tweets from them is not easy and the filtering result might be unsatisfactory. Hence, to collect news tweets effectively, we use several Twitter news accounts instead of selecting news tweets from general tweets. By using Twitter news accounts, we can collect news tweets without any complicated steps for extracting news tweets from the tweet collection. In this work, we considered Twitter news accounts in the USA according to the number of their followers. Although specific news accounts have many followers, we excluded Twitter news accounts covering specific topics such as politics, sports, travel, weather, etc., or providing only news video. As a result, we chose the top 20 Twitter news accounts (@nytimes, @CNN, @Reuters, @WSJ, @TIME, @FoxNews, @AP, @HuffingtonPost, @ABC, @washingtonpost, @NPR, @CBSNews, @NBCNews, @business, @Newsweek, @USATODAY, @Slate, @TheAtlantic, @NewsHour, and @VOANews). We collect the news tweet using the Twitter API and then parse the JSON format and store it.

We preprocess the collected tweets for more accurate news extraction. Even though typical news tweets contain a URL link to the website where the original news content is available, there are two exceptional cases. The first is when the news tweet does not include any URL. In this case, we just disregard it because its original content is not available. The second is when the URL in the tweet indicates a news video resource. In this case, we cannot collect news contents in text form and, therefore, we exclude these tweets, too.



Reuters Top News @Reuters · 11h
Germany produced record 35 percent of power from renewables in first half reut.rs/2sfEQkB pic.twitter.com/fLJg8QUAzM

CNN @CNN · 2h
A new study finds the numbers of Americans being diagnosed with opioid addiction climbed nearly 500% over 7 years cnn.it/2swoUOO

Desired Readability: 10
Desired Topic: Electric vehicle
Desired Popularity: Very Popular
Desired Period: Recently 1 month

Figure 1. Overall architecture of our proposed method.

3.3.2. Educational Video Collection and Preprocessing

We collect various multimedia data related to the news so that learners can easily understand the contents of the news. For this purpose, we use educational videos that are known to be more effective in learning compared to other media such as photographs and audio. Many educational videos are already available on the web. However, we use the TED talk video since is intended for educational purposes, very popular currently, and open to the public.

TED talk video has various metadata including speaker, duration, upload time, and script. In particular, script is very effective to infer the content without watching the whole video, and can be used to link with news depending on the purpose. Unfortunately, in TED, script data is not published in a textual form and includes some extra information. This extra information should be handled appropriately to acquire the script. For example, it is necessary to remove exclamations or stage descriptions [4]. For linking, we usually calculate the similarity between the script and the news. More details can be found in our previous work [4].

3.3.3. Physical Device Data Collection

Many diverse devices can be used in an IoT-based class. The most commonly used devices include Raspberry Pi, Arduino, and mobile devices. The learners generate various kinds of data as they use these devices in class. Among various data that can be generated from the devices in class, we used learners' interaction data on the Raspberry Pi and mobile device in the previous study [4]. In particular, we used Raspberry Pi to figure out the learner's level and to provide the learner with appropriate learning materials. In addition, learners could search for diverse data by formulating queries on the mobile device.

3.4. TNA (Twitter News Analyzer)

For the N-STEAM education, we need to extract only science/technology news. To do this, we classify the collected news tweets into two categories using various machine learning methods. Then, we analyze topics and make clusters by using a topic modeling method for the classified science/technology news tweets. In addition, we measure the readability level of the news to provide appropriate news to the learner for his/her level. Finally, we remove redundant news in the same cluster to create a more concise news data set.

3.4.1. Feature Term Selection

Since news tweets have a limited number of characters, proper feature term selection is required to effectively classify them [34]. A commonly used method for effectively and simply representing the features of text documents such as news tweets is the bag-of-words method. This method represents each document as a point in a vector space with one dimension for every term in the vocabulary [38]. A news tweet $t_i \in$ T can be represented by a vector $t = (\alpha_1, \alpha_2, \cdots, \alpha_m)$, where α_j is the frequency of a word j in t_i and m denotes the total number of words in t_i [17].

To obtain an optimal collection of terms from each news tweet for machine learning, we perform the following steps. Through these steps, we can reduce the feature space of the tweet corpus and, hence, achieve a more accurate classification performance.

1. Tokenization: Given a raw tweet, tokenization is the task of chopping it into pieces, called tokens, perhaps, at the same time, throwing away certain characters, such as punctuations. The major idea of tokenization is to break the entire document into a list of sentences [39,40]. Generally, tokens are separated by whitespace characters, such as a space and a line break, or by punctuation characters [39]. A wide variety of tokenizers have been developed, but we utilize the Natural Language Toolkit (NLTK) tweet tokenizer (http://www.nltk.org/api/nltk.tokenize.html), which is best suited for tweet processing [23]. Unlike other tokenizers, it has features suitable for tweet processing. That is, it tokenizes the URLs contained in the tweet as a single token to make it easier to deal with later. It also removes unnecessary words by removing usernames starting with "@". In this case, each tokenized word is called a term.
2. Dropping common terms (stop words): For each selected term, we count how many times the term appears in a tweet. However, some extremely common words could be of little value in the document selection. They should be excluded from the vocabulary entirely. These words are called stop words [40]. There are many stop word sets currently published. We first remove common English stop words (e.g., "a", "the", "and", "as", etc.). However, this is not enough to process Twitter data. This is because Twitter uses various special characters and abbreviations that

are not used in the past (e.g., "#", "RT (Retweet),""@", etc.). Therefore, we create a domain-specific stop word list suitable for news tweet processing [23]. For example, we remove the most frequently appearing terms such as "news", "says", etc., and the name of each news channel (e.g., "Reuters", "Nytimes", "CNN", "Foxnews", etc.) in a news tweet. We also consider "RT", URL, etc., as stop words and remove them from the news tweet. On the other hand, we do not remove the hyphen, which combines two words. For example, words such as "self-driving", "start-up", "high-tech", "e-commerce", "solar-powered", and "Wi-Fi" are meaningless when they are separated. We do not remove numbers either because numbers in science/technology news about electronic products, such as iPhone 7 and Galaxy Note 7, may refer to the version number of the product. In addition, since official news Twitter accounts do not use slang, we do not need to deal with them. Finally, we do not remove the hashtags. This is because they are usually created by a user to represent the theme or topic of the message.

3. Rare term removal: Term frequency is one of the key criteria in determining the importance of a term in a document. Since typical tweets can be in a mix of lowercase and uppercase letters, we consider every term of a document in lowercase and count its frequency. If the frequency is lower than some threshold, we consider the term as a rare term and ignore it. The reason we ignore rare terms is that if too many terms are used as features for the classification, the accuracy would deteriorate.
4. Stemming: Typical documents contain different forms of a word for grammatical reasons, such as "organize", "organizes", and "organizing". In addition, there are families of derivationally related words with similar meanings, such as "democracy", "democratic", and "democratization" [40]. Stemming is a procedure for reducing all words with the same stem to a common form. It is used to remove derivational suffixes as well as inflections so that word variants can be conflated into the same roots or stems [23]. Thus far, various stemming algorithms have been developed [41], the most popular of which include Porter's stemmer, Snowball (Porter II) stemmer, Lancaster stemmer, and Lovin's stemmer. In this paper, we will consider all these stemmers and use the most accurate one.
5. POS (Part-of-Speech) tagger: Since the selected feature terms have a wide variety of parts of speech, it is necessary to choose whether to use the words of all parts of speech. By extracting only nouns in most documents, we can easily deduce the meaning of the document. Therefore, we use a POS (Part-of-Speech) tagger. A POS tagger is a piece of software that reads text in some language and assigns a part of speech to each word (and other tokens), such as noun, verb, adjective, etc. We use the Stanford POS tagger (http://nlp.stanford.edu/software/tagger.shtml) to extract only nouns from news tweets and use them as feature sets.

3.4.2. News Tweet Classification

Since we are interested in science/technology news only, we need to perform filtering for news of diverse subjects. For the news tweet classification, we need to collect training data sets. We use Reuters' news Twitter account and website to collect the news tweets and dereference them using the URL in the collected tweets to parse the data of the original news article. If we parse an original news article, the category of the tweet can be found easily [23].

However, since many news websites use different languages and document structures, it is very time-consuming to parse the page structure and extract the relevant news according to the structure parsing. One possible solution to this problem is to determine the category of the news according to its content using a machine learning technique. To this end, we collect category information for Reuters news tweets according to the Reuters category criteria and use them as the training data set for the classifier [23].

Now, users can easily classify tweets related to science/technology simply by putting tweets collected from various tweet accounts into the classifier. We put all the news other than science/technology news into the "others" category for faster and more accurate classification [23].

3.4.3. News Topic Modeling

Once the science/technology news tweets are collected using the news tweet classifier, they need to be classified again by their topic for N-STEAM education [23]. Classification by topic can be done using the unsupervised methods such as k-means clustering. Those methods are usually good for classifying data without labels, but they have a limitation in that it is difficult to judge the semantic relationship between words. On the other hand, a topic modeling method is known to complement this limitation and can be used to effectively identify the semantic relationship between words. Hence, for grouping news tweets by topic, we use the Latent Dirichlet allocation (LDA) method, one of the most popular topic modeling methods. LDA is a generative probabilistic model for a collection of discrete data such as text corpora. With respect to text modeling, each document may be viewed as a mixture of various topics, where each topic is characterized by a distribution over words [42,43].

To use the LDA method for classification, we first dereference through the URL link in the science/technology news tweet to extract the original news text [23]. Since Twitter is mostly used to share specific content/sites or spread news, most of our news tweets have possibly multiple external URL links. If the original news content is available through the URL link, topic modeling can be done easily. This technique is called dereferencing. For example, Table 1 shows a tweet that contains two external links. The first link (https://t.co/y0vtZ6JUvQ) is to the original news, and the second link (https://t.co/4ydC86R3ys) is to the original tweet. Therefore, if we dereference through the first link in the tweet, we can collect the original news text for which we can perform topic extraction.

Table 1. Example of a dereferencing method.

LG Electronics says to invest in robot technology **https://t.co/y0vtZ6JUvQ**, https://t.co/4ydC86R3ys
↓
South Korea's LG Electronics Inc said on Sunday it will aggressively invest in robots, seeking to capitalize on advancing artificial intelligence that may eventually lead to sophisticated machines performing everyday human tasks. (Interruption . . .)

After the original news is collected by the dereferencing method, preprocessing (tokenizing, stop words elimination, stemming, etc.) is performed in a similar manner as described before. Finally, we will apply the LDA method to the news collection represented by the refined terms to see all the topics the news collection consists of [23].

3.4.4. Readability Analysis

In a news-based class, learners might lose interest and get bored if the news contents provided are too difficult or too easy. That is, providing appropriate news contents according to the learner's level is very important in the effectiveness of a news-based class. Here, readability represents an index that indicates how easily the reader can understand specific content and how readable it is [23]. To date, many readability metrics have been proposed including Coleman–Liau [44] and Flesch Reading Ease [45]. Some of the popular metrics can be defined as follows:

- Coleman–Liau Index

$$\alpha_1 = 0.0588L - 0.296S - 15.8 \tag{1}$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

- Flesch Reading Ease

$$\alpha_2 = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \tag{2}$$

- Flesch–Kincaid Grade Level

$$\alpha_3 = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59 \tag{3}$$

- Gunning Fog Index

$$\alpha_4 = 0.4\left[\left(\frac{words}{sentences}\right) + 100\left(\frac{complex\ words}{words}\right)\right] \tag{4}$$

- SMOG

$$\alpha_5 = 1.0430\sqrt{Number\ of\ polysyllables \times \frac{30}{Number\ of\ sentence}} + 3.1291 \tag{5}$$

- Automated Readability Index (ARI)

$$\alpha_6 = 4.71\left(\frac{characters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43 \tag{6}$$

These metrics except the Flesch Reading Ease give the readability level of a given news item in the range of 1 to 22 (US grade levels) [4,23]. Since the Flesch Reading Ease metric gives the readability in the range of 100 to 1, we normalize the metric into a scale of 1 to 22 using the following equation [4]:

$$\alpha_{2.1} = \begin{cases} 22 - 2{\cdot}\alpha_2 - 2{\cdot}\frac{\alpha_2\ \%\ 10}{10}\ \left(0 \le \frac{\alpha_2}{10} < 7\right) \\ 15 - \frac{\alpha_2}{10} - \frac{\alpha_2\ \%\ 10}{10}\ \left(7 \le \frac{\alpha_2}{10}\right) \end{cases} \tag{7}$$

A comparison of these metrics showed that ARI tends to give scores that are higher than those of Kincaid and Coleman–Liau, but are usually slightly lower than those of Flesch. Moreover, the Kincaid formula is probably the best predictor for technical documents. Both ARI and Flesch tend to overestimate the difficulty, whereas Coleman–Liau tends to underestimate it [46]. Owing to this property, to calculate the readability level of the news, we do not rely on a specific metric but combined all the metrics using (8). We also consider the final readability level of each news tweet as one of its feature values so that we can easily find out the right news contents for the learner's level.

$$\alpha_{total} = \frac{\alpha_1 + \alpha_{2.1} + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6}{6} \tag{8}$$

3.4.5. News Filtering

Considering the growth rate of SNS data, the amount of news that can be collected through various news accounts is enormous. Moreover, it is likely that there are multiple news articles on the same topic even though their readability could be different. Therefore, for effective news browsing, such duplicated news articles need to be detected and removed. We remove such duplicate news articles into two stages. In the first stage, we calculate the similarity of all the news classified into a specific topic. To compare the similarity between the words in the news, we use the cosine similarity method. Therefore, if the cosine similarity between two news articles is greater than some threshold, then we consider them as duplicate news. In the second stage, we consider the readability of all the duplicated news articles in terms of user level. If there are two or more duplicate news articles, we keep the news that has a lower or the lowest readability score and remove the rest. For example, if there are six duplicate news articles and their readability levels are 9.1, 9.4, 9.6, 10.2, 11.4, and 11.7, respectively, then we remove the news with 9.1 and remove the other news (9.4 and 9.6) for the 9th grade. Similarly, we keep the news with 10.2 because that is the only one for the 10th grade. Finally, we keep the news with 11.4 and remove the news with 11.7. As a result, after removing the duplicates, we have three news articles whose readability levels are 9.1, 10.2, and 11.7, respectively. Through this, we are given a set of news articles with minimal duplication. Finally, we rank the news tweets by considering the number of favorites and retweets in the news tweets.

3.5. Ontology Component

Since the amount of collected data is very large, learners have difficulty in selecting appropriate materials in the N-STEAM lessons. To solve this problem, we produce semantic annotations for the Twitter news and related materials. In this section, we describe how to develop an LNT (Linked News Tweet) ontology to represent the collected Twitter news and diverse metadata, and produce semantic annotation using the ontology.

3.5.1. DBpedia Entity Extraction

In DBpedia, structured content is extracted from the information in Wikipedia. This structured content is made available on the web. DBpedia allows users to semantically query the relationships and properties of Wikipedia resources, including the links to other related data sets. We enrich the semantic information of news documents using DBpedia. To do this, we first fetch the original news text via the URL in the collected Twitter news. Second, to extract the Wikipedia concept (entity), we perform a concept matching on the extracted news text using DBpedia Spotlight (https://github.com/dbpedia-spotlight/dbpedia-spotlight). This allows access to a vast number of established taxonomies and vocabularies such as DBpedia. Through this, unstructured free text is enriched with the unique Uniform Resource Identifiers (URIs) of structured Linked Data entities to allow further reasoning on related concepts and enable learners to query for resources using well-defined concepts and terms [23,47]. Table 2 shows some examples of DBpedia concept extraction.

Table 2. Examples of DBpedia concept extraction.

Type	Content
Surface Form(concept)	Tokyo
URI	http://dbpedia.org/resource/Tokyo
Concept Type	Schema:Place,DBpedia:Place,DBpedia:PopulatedPlace, DBpedia:Settlement,Schema:City,DBpedia:City
...	...

3.5.2. Semantic Annotation

Semantic annotation uses ontology objects to enrich a resource's information that tells a computer the meanings and relations of the data terms. In other words, it is the process of linking an electronic resource to a specific ontology. Electronic resources can be text contents, images, video, services, etc. [48].

For the semantic annotation, we first design an appropriate ontology (LNT ontology) by reusing an existing ontology language [48–50]. More specifically, Twitter news is represented as a microblog post in the sense of the SIOC (http://sioc-project.org/) (Semantically Interlinked Online Communities) ontology [50]. The SIOC ontology aims to enable the integration of online community information. The SIOC provides a semantic web ontology for representing rich data from the social web in RDF. It is now a standard vocabulary for expressing social data in RDF [18]. The maker of the tweet is represented using the FOAF (http://www.foaf-project.org/) (Friend of a Friend) ontology [50]. FOAF is a project devoted to linking people and information using the web. It is used to represent users, as it provides a simple way to describe people, their main attributes, and their social acquaintances [18]. General properties are represented using the Dublin Core vocabulary (http://dublincore.org/specifications/) [50]. It is a lightweight RDFS (RDF Schema) vocabulary for describing generic metadata. The combination of these ontologies forms a complete structure that represents the Twitter news and various pieces of relevant information.

Now, we design the ontology schema using the terms mentioned above. An ontology is a model for describing a world that consists of a set of types, properties, and relationship types. An ontology consists of classes, instances (individuals), relations, and properties [51]. For the ontology schema design, we first create the classes and determine their relationships. We also create several types of properties to determine the relationship between resources. Tables 3–5 show the class, object property, and datatype property, respectively, used in our ontology. We also use common RDF/XML data formats for this modeling to provide easy reuse across semantic web-based applications, notably by using SPARQL for querying [13].

For the ontology instance creation, we collect Twitter news, related information including DBpedia entity, physical device data, and news company information from the influence tracker website [52]. Razis and Anagnostopoulos [52] claimed that it is difficult to judge the influence of a Twitter account just by using the number of followers because there are many other factors such as the number of followers, the retweet percentage, the tweet per day, etc. Thus, we add the metadata presented in the influence tracker to the ontology schema and create its instances. We then upload other collected data into the instance according to our ontology schema.

Table 3. Lnt ontology class.

Name	Description
lnt:PhysicalDevice	A class that stores data generated by physical devices.
sioc:Post	A class that stores tweets and related information.
sioc:UserAccount	A class that stores Twitter user accounts and related information.
sioc:Topic	A class that stores topics for each news.
lnt:Concept	A class that stores concepts and related URIs obtained through DBpedia Spotlight.
lnt:ConceptTypeURI	A class that stores information related to concept types.
lnt:ConceptURI	A class that stores concept-related information.
lnt:EduVideo	A class that stores educational video material.

Table 4. Lnt ontology object property.

Name	Description
lnt:hasPhysicalDevice	A property to connect from Post class to PhysicalDevice class.
lnt:hasAnnotation	A property to connect from Post class to Concept class.
lnt:hasConceptURI	A property to connect from Concept class to ConceptURI class.
lnt:hasTypeURI	A property to connect from Concept class to ConceptTypeURI class.
sioc:has_topic	A property to connect from Post class to Topic class.
sioc:has_creator	A property to connect from Post class to UserAccount class.
lnt:hasEduVideo	A property to connect from Post class to EduVideo class.

Table 5. Lnt ontology datatype property.

Name	Description
lnt:sensor	A property that stores various sensor data.
lnt:raspberryPi	A property that stores the Raspberry Pi data.
lnt:arduino	A property that stores the Arduino data.
lnt:mobileDevice	A property that stores the mobile device data.
dcterms:created	A property that stores the date and time when the news tweet was created.
dcterms:title	A property that stores the title of the original news obtained by dereferencing the URI of the news tweet.
foaf:homepage	A property that stores the homepage of the news company.
foaf:name	A property that stores the name of the news company.
sioc:content	A property that stores the plain-text of the original news.
sioc:id	A property that stores the user ID of the one who posted the news tweet.
lnt:topic_num	A property that stores the number of the classified topic using a topic modeling method.
lnt:topic_word	A property that stores the words used in a specific topic.
lnt:tweetNum	A property that stores the number of tweets we saved.
lnt:favoriteCount	A property that stores the number of favorites of a specific news tweet.
lnt:retweetCount	A property that stores the number of retweets in a specific news tweet.
lnt:tweetURI	A property that stores the URI that is linked to the news tweet.
lnt:newsURI	A property that represents the URI that is linked to the original news text.
lnt:tweetContent	A property that saves the contents of a plain-text message with the URI removed from the news tweet.
lnt:readability	A property that stores the readability level of the original news text.
lnt:retweetPercentage	A property that indicates the average percentage at which a specific news tweet is retweeted.
lnt:tweetPerDay	A property that indicates the average number of news tweets posted by a specific Twitter account per day.
lnt:followingCount	A property that indicates the number of other accounts that the Twitter news account follows.
lnt:followerCount	A property that indicates the number of users following a specific Twitter account.
lnt:tweetCount	A property that indicates the total number of news tweets posted by a specific Twitter news account.
lnt:influence	A property that stores the influence index of a Twitter news account.
lnt:conceptType	A property that stores the type of extracted DBpedia concepts.
lnt:typeURI	A property that stores the URI of the conceptType.
lnt:conceptName	A property that stores the concept names
lnt:conceptURI	A property that stores the URI of extracted DBpedia concepts.
lnt:videoURI	A property that stores the URI of the educational video.
lnt:videoScript	A property that stores the script of the educational video.

3.5.3. RDF Triple Store

Linked-Data-based systems usually build on triple stores as their main data storage. This triple-based representation enables the integration of data available from various sources without the need for physical storage of the RDF triples that correspond to the relational data [53]. These systems provide data management and data access via application programming interfaces (APIs) and query languages to RDF data. For this work, we used the Jena Fuseki triple store (http://jena.apache.org/documentation/serving_data/). It provides representational state transfer (REST)-style SPARQL HTTP update, SPARQL query, and SPARQL update using the SPARQL protocol over HTTP [54,55]. In other words, we put the instance into our ontology schema, upload it to the Jena Fuseki triple store, and then the user gets the desired data through a SPARQL query.

4. Experimental Results

4.1. News Tweet Classification

To evaluate the performance of our scheme, we implemented a prototype system and performed several experiments. We collected 6488 science/technology news tweets and 28,504 other-category news tweets. We applied several feature selection methods described above to change the type of the feature set and then compared the accuracy of those classifiers. Figure 2 shows all the steps for the feature set construction.

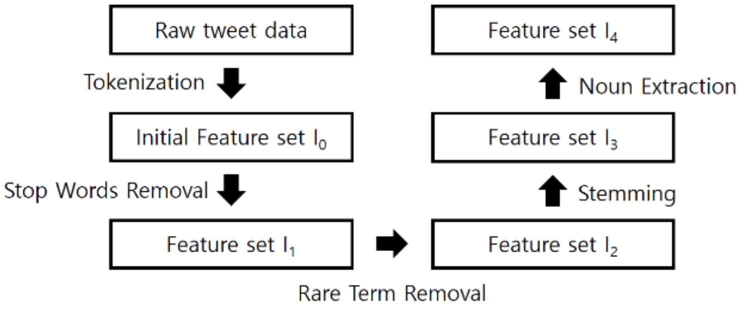


Figure 2. Feature set construction process.

The classification accuracy is one of the indicators that can be used to evaluate the performance of the classifier and refers to the rate at which it is correctly classified when the test set is given to the trained classifier. Moreover, the data included in the training and test sets were shuffled each time to prevent bias. Experimental results were obtained by performing five measurements using a shuffled data set and averaging the values. The classifiers that we considered included Adaboost (1), Naïve Bayes (2), multinomial Naïve Bayes (3), Bernoulli Naïve Bayes (4), LinearSVM (5), nu-SVM (6), logistic regression (7), and stochastic gradient descent (8), which have been widely used. Table 6 shows our hypotheses and experimental results. In the first experiment, we assumed that we could achieve faster and more accurate classification when we used the terms of science/technology news tweets only. For comparison, we investigated the outcome of news tweet classification when using all terms and using only science/technology terms. The experimental results are shown in Tables 6 and 7, respectively. The results showed that most of the classifiers showed similar or slightly higher accuracy when using all terms, as shown in Table 6. In other words, it would be better to use all terms in the data set to improve the classification accuracy of news tweets. However, classification accuracy was not much different when using the science/technology terms only, as shown in Table 7. However, the classification was faster since we used a much smaller set of terms.

In the second experiment, we investigated the effect of removing rare terms on the classification accuracy of news tweets. To remove the rare terms, we first counted the number of terms appearing in the entire data set. Then, the terms with the number of occurrences less than a predefined threshold were removed for the classification. Experimental results showed that the accuracy of the classification was slightly improved ($I_{2\text{-}1}$, $I_{2\text{-}2}$) when the terms appearing once or twice in the total data set were removed. Among them, the set $I_{2\text{-}2}$ could improve the accuracy while utilizing a fewer number of terms and was the most efficient feature set.

In the third experiment, we investigated the performance of various stemmers on the classification accuracy. To see the effect of the stemmers, we used the feature set $I_{2\text{-}2}$, which was found to be the most efficient. The result showed that the Porter stemmer did not show any significant difference in the classification accuracy whether it was used or not. However, the Lancaster stemmer and Snowball stemmer showed a significant improvement for all the classifiers. Overall, the Snowball stemmer showed better accuracy than the Lancaster stemmer even though the difference was very little.

Usually, nouns are the main parts of speech that represent the core contents. If we ignore other parts of speech and use nouns only for the classification, we could still get a reasonable performance, while reducing the number of terms for the classification. Thus, in the fourth experiment, we just considered nouns in the news tweets (I_4) and measured the classification accuracy. As a result, compared with the set $I_{3\text{-snowball}}$, most of the classifiers showed a similar performance for I_4. This indicates that we could get a similar classification accuracy in less time by using a much smaller set of terms.

In the fifth experiment, we evaluated the effect of the ratio of the training set to the test set on the classification accuracy. The ratios of the training set to the test set we considered in this experiment were 5:5, 6:4, 7:3, 8:2, and 9:1. To reduce the variability of the data set, we performed multiple rounds of cross-validation using different partitions and averaged the validation results over the rounds. We also used the feature set $I_{3\text{-snowball}}$ for this experiment. Table 8 shows the result. As shown in the table, the overall accuracy was best when the ratio was 8:2. Hence, we performed all the classifications using this ratio.

Table 6. News tweet classification results (all terms).

Feature Set	No. of Terms	Classifier							
		1	2	3	4	5	6	7	8
I_0	17,121	78.91	85.03	84.88	81.92	85.59	83.31	86.25	84.07
I_1	16,959	79.07	84.97	85.18	81.23	85.43	82.96	85.81	85.29
$I_{2\text{-}1\ (=1)}$	10,842	78.95	85.55	85.55	84.37	85.45	82.62	86.02	85.48
$I_{2\text{-}2\ (\leq 2)}$	8018	79.25	85.67	85.74	85.53	85.10	82.94	85.98	85.18
$I_{2\text{-}3\ (\leq 3)}$	6378	78.95	85.58	85.70	85.54	84.88	82.77	85.84	85.05
$I_{3\text{-porter}}$	6184	79.92	85.30	85.50	85.37	85.04	82.75	85.81	84.70
$I_{3\text{-lancaster}}$	5536	86.57	91.51	91.31	91.58	91.55	87.10	92.31	91.44
$I_{3\text{-snowball}}$	5988	87.71	91.69	91.61	91.93	92.44	87.70	93.06	92.47
I_4	3662	87.52	91.61	91.49	91.63	91.99	87.14	92.65	92.26

Table 7. News tweet classification results (science/technology terms).

Feature Set	No. of Terms	Classifier							
		1	2	3	4	5	6	7	8
I_0	7486	78.72	85.77	84.87	84.26	85.10	82.18	85.02	84.03
I_1	7346	78.77	85.82	84.59	84.16	84.89	82.03	84.91	84.62
$I_{2\text{-}1\ (=1)}$	4712	79.18	85.77	84.37	85.08	85.22	81.83	84.98	84.98
$I_{2\text{-}2\ (\leq 2)}$	3421	79.00	85.44	84.12	85.22	85.15	82.31	84.96	84.85
$I_{2\text{-}3\ (\leq 3)}$	2633	79.41	85.06	83.80	84.80	85.02	82.16	84.63	84.56
$I_{3\text{-porter}}$	2802	79.63	84.98	84.36	84.95	84.60	81.35	84.64	83.98
$I_{3\text{-lancaster}}$	2611	86.01	91.06	90.55	91.16	91.19	86.46	91.36	90.91
$I_{3\text{-snowball}}$	2729	87.29	91.68	90.98	91.78	92.04	86.41	92.14	91.91
I_4	1585	87.67	91.10	90.36	91.25	92.13	86.89	91.67	91.59

Table 8. Classification accuracy according to training/test set ratio change.

Training/Test	1	2	3	4	5	6	7	8
5:5	87.05	90.62	90.67	90.95	91.10	87.57	91.50	90.55
6:4	87.99	91.26	91.16	91.27	91.63	87.20	92.10	91.06
7:3	87.90	91.39	91.33	91.55	92.05	87.51	92.67	91.80
8:2	87.73	92.05	91.85	92.24	92.39	88.27	93.26	92.67
9:1	87.59	91.71	91.61	91.87	92.67	88.06	92.36	92.19

From these experiments, we could see that the classification performance was improved as the feature set was refined from I_0 to I_4.

4.2. News Topic Modeling

In this section, we show how to extract topics from the collection of science/technology news using the LDA method. When a document set is modeled with an optimal number of topics, the likelihood becomes largest. The optimal number of topics for a document set can be found easily by checking the likelihood for varying the number of topics (k) [23]. To determine the optimal number of iterations for the LDA model, we first modeled topics by changing the number of topics (k) from 10 to 100 (increasing by 10). Experimental results showed that most likelihood values were stabilized at the point where the number of iterations was about 4000 (Figure 3). Thus, we did 4000 iterations to find out the most appropriate number of topics.

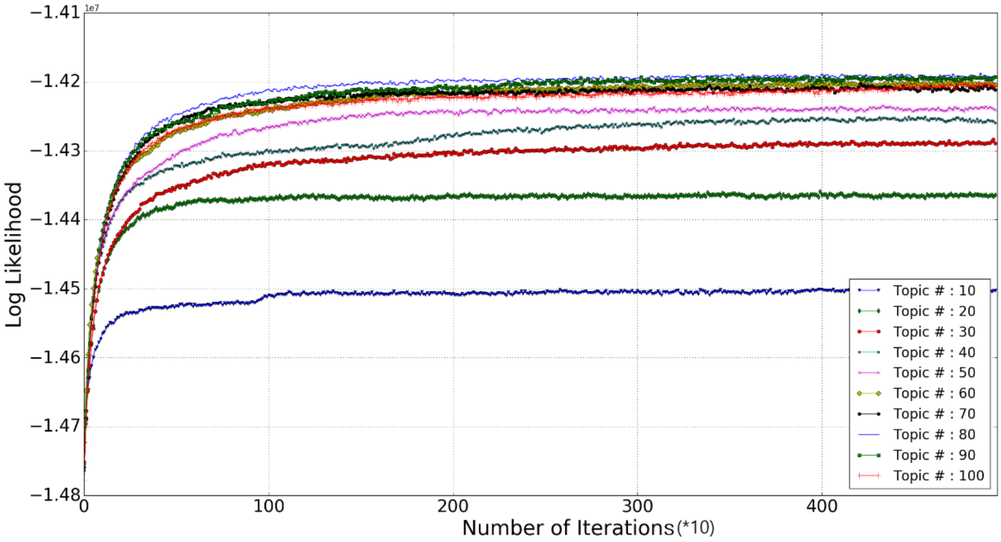


Figure 3. Experimental results for the iteration stabilization point.

Second, we changed the value of k from 1 to 200 (increasing by 5). To find the best number of topics in our data set (Figure 4.), we checked the k value with the highest likelihood value. The log-likelihood value increased rapidly when k was 5 to 30 and gradually decreased when k was greater than 80. In addition, the highest log-likelihood value was obtained when k was 80; a relatively high likelihood value was also obtained when k was 70–100. Therefore, we performed the topic modeling by setting k to 80, and an efficient topic modeling could be possible even when k was 70–100.

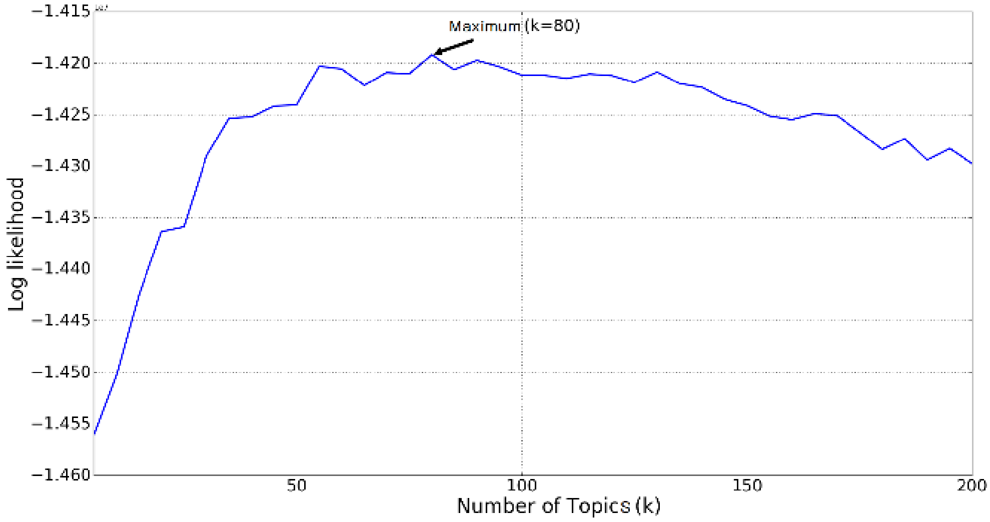


Figure 4. Changes in log likelihood with changes in topic count.

Third, we used the chosen k value to see what topics were present in our data set. Table 9 shows the result. For instance, the first topic could be space technology considering the top-level words such as “mars”, “planet”, “earth”, etc. The second topic could be Amazon’s delivery of goods using drones considering the words such as “amazon”, “delivery”, “drone”, etc. In a similar way, the third topic could be an electric car. The other topics can be inferred by using the extracted top-level words. Through this experiment, we could see that the words of each topic extracted by the LDA method represented the topic quite clearly. Overall, we can conclude that Twitter news data can be used effectively to provide learners with appropriate and diverse contents of science and technology topics.

Table 9. Top-level words for each topic.

Topic ID	Top-Level Words
1	mars, planet, said, space, mission, earth, scientists, system, surface, life, solar, billion, european . . .
2	amazon, delivery, said, drones, online, food, service, new, would, drone, retailer, inc, customers, stores . . .
3	electric, cars, car, vehicles, said, battery, new, vehicle, motor, sales, bmw, volkswagen, toyota, vw . . .
4	space, station, nasa, said, glenn, astronauts, first, moon, earth, mission, flight, program, orbit, astronaut . . .
5	flight, airlines, said, aircraft, solar, plane, flights, airline, air, seats, powered, boeing, seat, four . . .
6	taiwan, said, sharp, display, co, year, robot, robots, foxconn, percent, inc, million, apple, ltd, billion . . .
7	high, first, project, power, technology, one, system, hyperloop, 5g, would, speed, company, network . . .
8	samsung, note, said, recall, phones, customers, fire, devices, galaxy, batteries, replacement, smartphone . . .
. . .	. . .
78	intel, chips, production, lenovo, pc, corp, 3d, company, used, also, sony, manufacturing, gopro . . .
79	google, reality, said, devices, virtual, device, new, assistant, technology, voice, vr, smart, home, users...
80	launch, spacex, said, rocket, space, satellite, musk, company, would, falcon, nasa, air, force, station . . .

Finally, we calculated the distribution of news tweets in the data set over 80 topics, as shown in Figure 5. In the figure, we can see that there are especially many news articles on topics 12 and 33, and that the overall distribution is various depending on the topics.

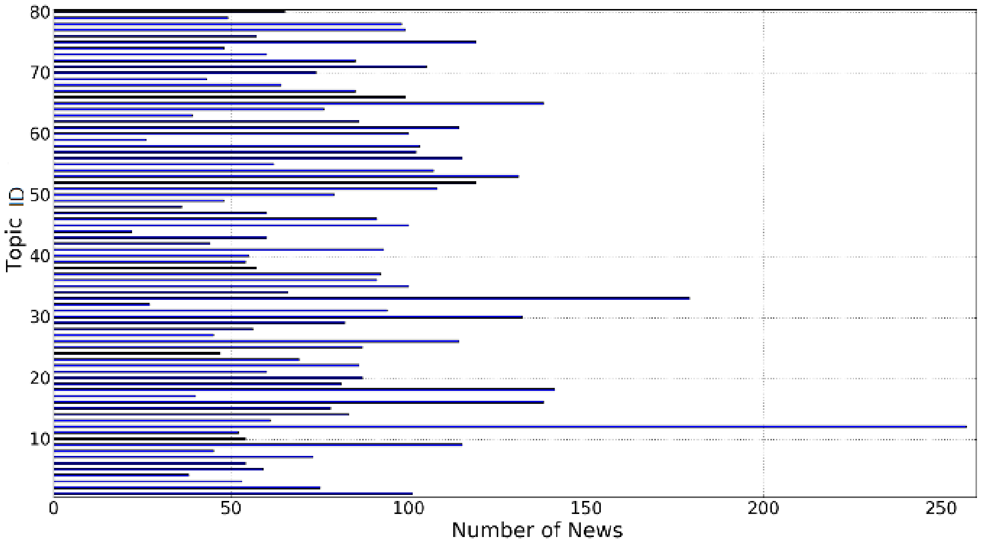


Figure 5. Distribution of news documents by topic.

4.3. News Readability Measurement

In this section, we describe the readability measurement on the original news extracted from news tweets. The readability level of the original news text was measured using six readability metrics, namely, Flesch–Kincaid (α_1), Flesch Reading Ease ($\alpha_{2.1}$), GunningFog (α_3), SMOG (α_4), ColemanLiau (α_5), and ARI (α_6).

Table 10 shows the result. Depending on the news contents, the metrics showed quite a significant difference in the readability grades. For this reason, we used their average value. Figure 6 shows some of the readability measurement statistics for the data set. According to the figure, our science/technology news articles had a wide variety of readability grades. The average grade of all the news articles was 11.86, which indicates that most of them could be read without difficulty if the learner had the reading level of high school students. In addition, the final readability grade for each news item in Figure 7 showed that most of these news articles were concentrated between grade 9 and grade 13.

Table 10. Readability measurement result.

News	Readability Metric						Average Grade	Final Grade
	α_1	$\alpha_{2.1}$	α_3	α_4	α_5	α_6		
#1	11.25	12.4	12.85	13.9	12.64	12.05	12.52	12
#2	10.76	11.6	12.95	13.85	11.96	11.71	12.14	12
#3	11.94	12.6	14.04	15.25	11.65	12.24	12.95	12
#4	10.85	11.8	12.99	14.61	12.45	11.95	12.44	12
#5	13.22	12.8	16.63	15.83	12.1	14.77	14.23	14
#6	10.85	11.8	12.99	14.61	12.45	11.95	12.44	12
#7	11.43	12	14.54	14.59	12.32	12.74	12.94	12
#8	13.22	12.8	16.63	15.83	12.1	14.77	14.23	14
#9	10.85	11.8	12.99	14.61	12.45	11.95	12.44	12
#10	9.91	11.4	11.07	12.82	11.35	10.05	11.1	11
...	...	...	...	...	...	...	...	...
Average	10.54	11.06	13.24	13.64	11.17	11.51	11.86	11.35

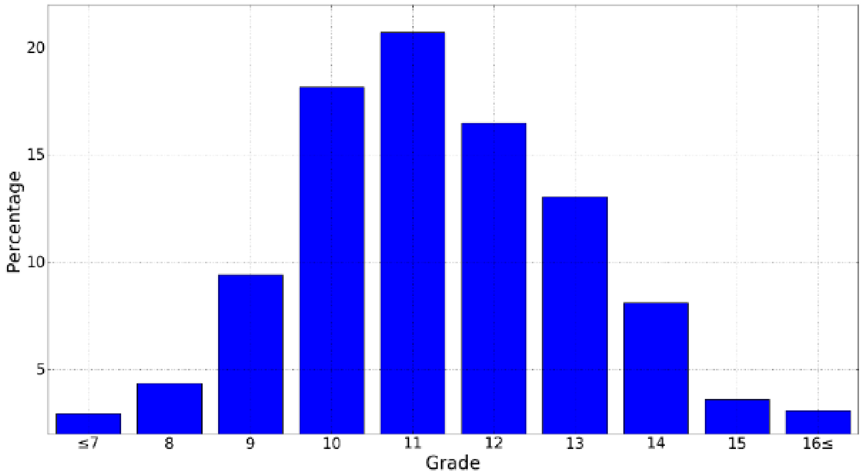


Figure 6. Readability measurement statistics.

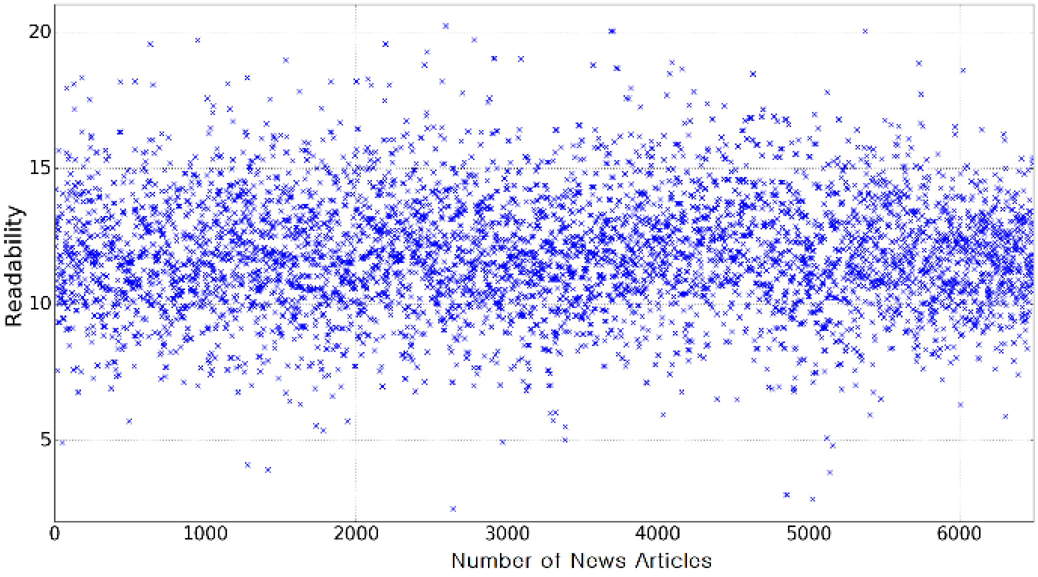


Figure 7. Readability measurement result.

In conclusion, for an effective N-STEAM education, it is necessary to select the appropriate news for the learners, depending on their learning ability. Another observation was that there are many diverse levels of news that can be used for middle and high school learners. Therefore, news contents are especially effective for the N-STEAM education of middle and high schools.

4.4. Semantic Annotation

In the final experiment, we designed the ontology schema and added instances to annotate the collected data. We also verified that RDF/XML files work well through various types of SPARQL queries. Through this experiment, we were able to confirm that our RDF/XML file can be used as a material for providing the latest science/technology contents that are effective for instructors and learners in N-STEAM education lessons. Finally, the RDF/XML file can also be used in a variety of places such as being published as Linked Data.

We reused a variety of existing ontology languages to build our LNT ontology. We used Protégé (http://protege.stanford.edu/), a popular ontology edition software, to construct and modify the ontologies. In addition, WebVOWL (http://vowl.visualdataweb.org/webvowl.html) was used for the effective visualization of the ontology schema. Figure 8 shows our ontology schema. In the figure, rectangles represent each class and the connections between the classes are represented using the object property. The datatype property is the content of each rectangle (class) [23].

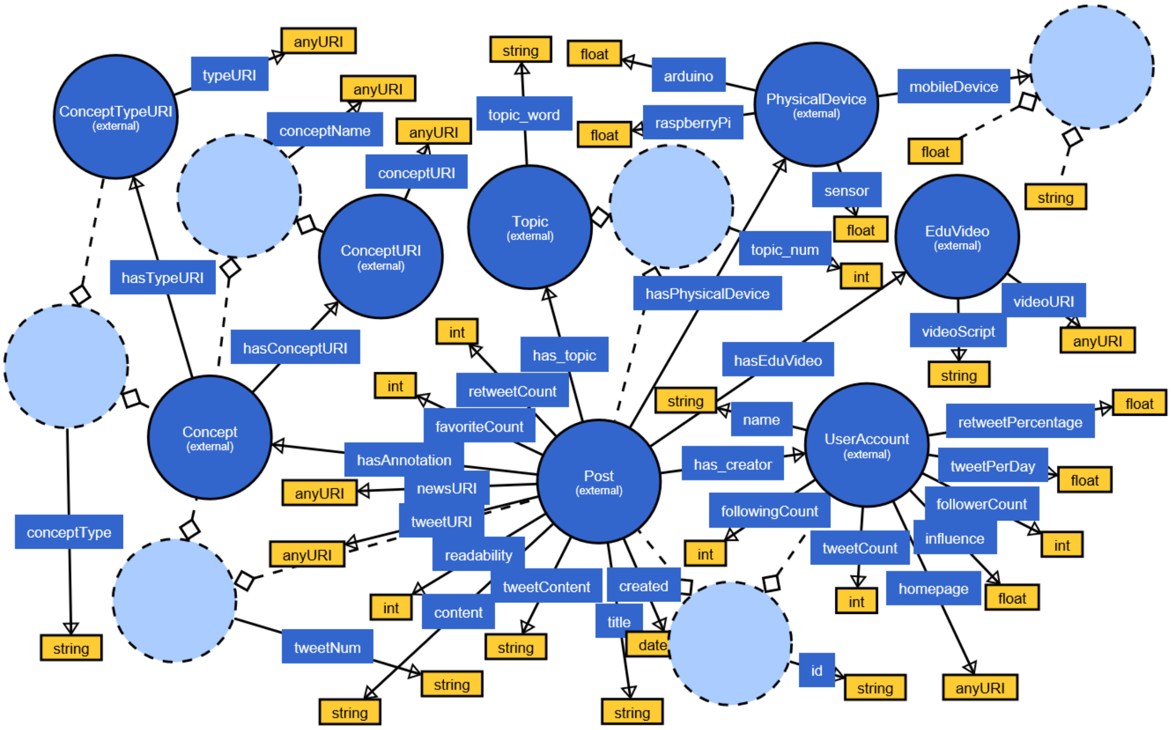
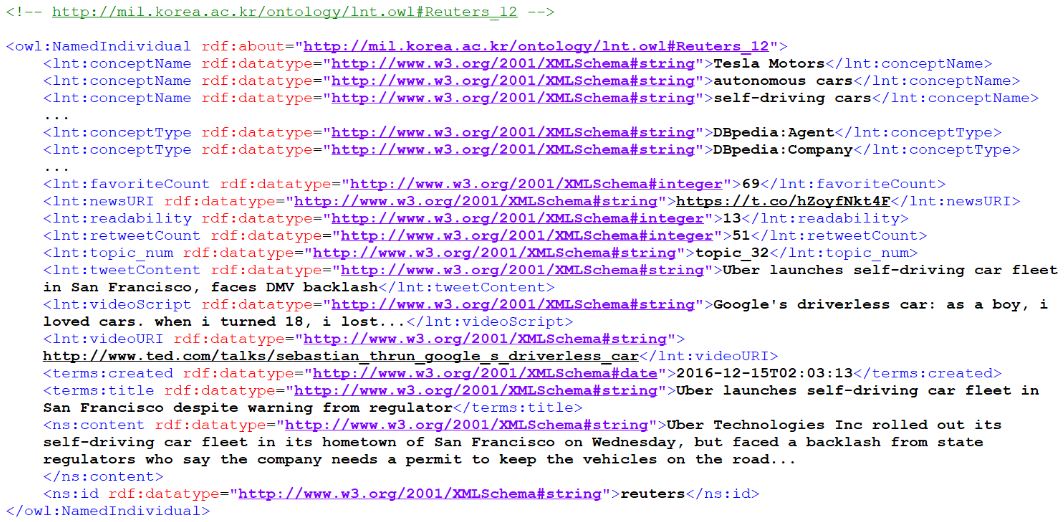


Figure 8. Visualization of Our Ontology.

Now, we show the result of annotating various data (Twitter news, Twitter news accounts, topics, DBpedia entities, etc.) as an RDF/XML file using an ontology (SIOC, FOAF, DC, and LNT). To this end, we collected the original data via the Twitter API, influence tracker, physical device, and DBpedia Spotlight. We also imported these data into the RDF/XML file as an instance and uploaded the final result to the web (http://mil.korea.ac.kr/ontology/lnt.zip) [23]. Instructors and learners can upload the final result file to Jena Fuseki and get a great deal of information through various SPARQL queries. An example of our final RDF/XML file is shown in Figure 9. We checked the number of triples uploaded in this file, and we identified a total of 311,413 triples.



```
<!-- http://mil.korea.ac.kr/ontology/lnt.owl#Reuters_12 -->

<owl:NamedIndividual rdf:about="http://mil.korea.ac.kr/ontology/lnt.owl#Reuters_12">
    <lnt:conceptName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Tesla Motors</lnt:conceptName>
    <lnt:conceptName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">autonomous cars</lnt:conceptName>
    <lnt:conceptName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">self-driving cars</lnt:conceptName>
    ...
    <lnt:conceptType rdf:datatype="http://www.w3.org/2001/XMLSchema#string">DBpedia:Agent</lnt:conceptType>
    <lnt:conceptType rdf:datatype="http://www.w3.org/2001/XMLSchema#string">DBpedia:Company</lnt:conceptType>
    ...
    <lnt:favoriteCount rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">69</lnt:favoriteCount>
    <lnt:newsURI rdf:datatype="http://www.w3.org/2001/XMLSchema#string">https://t.co/hZoyfNkt4F</lnt:newsURI>
    <lnt:readability rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">13</lnt:readability>
    <lnt:retweetCount rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">51</lnt:retweetCount>
    <lnt:topic_num rdf:datatype="http://www.w3.org/2001/XMLSchema#string">topic_32</lnt:topic_num>
    <lnt:tweetContent rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Uber launches self-driving car fleet
    in San Francisco, faces DMV backlash</lnt:tweetContent>
    <lnt:videoScript rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Google's driverless car: as a boy, i
    loved cars. when i turned 18, i lost...</lnt:videoScript>
    <lnt:videoURI rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    http://www.ted.com/talks/sebastian_thrun_google_s_driverless_car</lnt:videoURI>
    <terms:created rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2016-12-15T02:03:13</terms:created>
    <terms:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Uber launches self-driving car fleet in
    San Francisco despite warning from regulator</terms:title>
    <ns:content rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Uber Technologies Inc rolled out its
    self-driving car fleet in its hometown of San Francisco on Wednesday, but faced a backlash from state
    regulators who say the company needs a permit to keep the vehicles on the road...
    </ns:content>
    <ns:id rdf:datatype="http://www.w3.org/2001/XMLSchema#string">reuters</ns:id>
</owl:NamedIndividual>
```

Figure 9. Example of an RDF document.

To verify our final result, we performed a SPARQL query on our RDF/XML file for a specific condition and verified the result. We queried for all classes, instances, and their properties and cross-checked the results. We also confirmed that all the identified instances were returned [42]. We performed various types of test queries and confirmed the results. Here is an example of our query: (1) Retrieve all Twitter news with a specific topic; (2) Retrieve news of a specific period that includes a specific DBpedia concept; (3) Retrieve the Twitter news posted from a news account with a specific influence index; (4) Retrieve a news tweet whose favorite and retweet times are more than a certain number of times. We expressed the above queries in the SPARQL querying language and executed via Jena Fuseki [56]. Through this, learners can easily select news for their level and search for influential news and similar news when doing N-STEAM education lessons. In addition, through a DBpedia entity linked to a news article, various external related resources can be identified and utilized.

5. Conclusions

In this paper, we proposed a semantic annotation scheme that can effectively manage multimedia contents for learning and user interaction data with devices in the educational system for sustainable computing in the IoT environment. To do that, we first showed how to classify the most popular science/technology news articles from many Twitter news sources. Then, we showed how to extract their topics and calculate their readability levels using well-known metrics. Finally, we showed how to create an RDF/XML document using our ontology. To show the feasibility of our method, we performed quite extensive experiments for Twitter news to show the effectiveness and feasibility of our scheme. The results showed that our scheme can be used effectively for differentiated N-STEAM education. Since our method is quite generic, it can be used to cover various other educational fields with minor adaptation. In future work, we will explore better methods to further improve learning efficiency using more diverse physical devices and multimedia data.

Acknowledgments: This work was partly supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. R0190-16-2012, High Performance Big Data Analytics Platform Performance Acceleration Technologies Development) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1A09919590).

Author Contributions: Yongsung Kim planned the idea and processed the data. Yongsung Kim then generated the results and wrote the papers. Jihoon Moon collected experimental data, processed the data and visualized the experimental results. Eenjun Hwang conceived and supervised the work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, S.; Xu, L.D.; Zhao, S. The internet of things: A survey. *Inf. Syst. Front.* **2015**, *17*, 243–259. [CrossRef]
2. Ashton, K. That 'internet of things' thing. *RFiD J.* **2011**, *22*, 97–114.
3. Whitmore, A.; Agarwal, A.; Xu, L.D. The internet of things—A survey of topics and trends. *Inf. Syst. Front.* **2015**, *17*, 261–274. [CrossRef]
4. Kim, Y.; Jung, S.; Ji, S.; Hwang, E.; Rho, S. Iot-based personalized nie content recommendation system. *Multimed. Tools Appl.* **2018**, 1–35. [CrossRef]
5. Gómez, J.; Huete, J.F.; Hoyos, O.; Perez, L.; Grigori, D. Interaction system based on internet of things as support for education. *Proced. Comput. Sci.* **2013**, *21*, 132–139. [CrossRef]
6. Atzori, L.; Iera, A.; Morabito, G. The internet of things: A survey. *Comput. Netw.* **2010**, *54*, 2787–2805. [CrossRef]
7. Gubbi, J.; Buyya, R.; Marusic, S.; Palaniswami, M. Internet of things (IoT): A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **2013**, *29*, 1645–1660. [CrossRef]
8. Elmisery, A.M.; Rho, S.; Botvich, D. A fog based middleware for automated compliance with OECD privacy principles in internet of healthcare things. *IEEE Access* **2016**, *4*, 8418–8441. [CrossRef]
9. Rathore, M.M.; Ahmad, A.; Paul, A.; Rho, S. Urban planning and building smart cities based on the internet of things using big data analytics. *Comput. Netw.* **2016**, *101*, 63–80. [CrossRef]
10. He, J.; Lo, D.C.-T.; Xie, Y.; Lartigue, J. In Integrating internet of things (iot) into stem undergraduate education: Case study of a modern technology infused courseware for embedded system course. In Proceedings of the Frontiers in Education Conference (FIE), Erie, PA, USA, 12–15 October 2016; pp. 1–9.
11. He, J.S.; Ji, S.; Bobbie, P.O. Internet of things (iot)-based learning framework to facilitate stem undergraduate education. In Proceedings of the SouthEast Conference, Kennesaw, GA, USA, 13–15 April 2017; ACM: New York, NY, USA; pp. 88–94.
12. Oren, E.; Möller, K.; Scerri, S.; Handschuh, S.; Sintek, M. What are semantic annotations. *Relat. Tec. DERI Galway* **2006**, *9*, 62.
13. Mendes, P.N.; Passant, A.; Kapanipathi, P.; Sheth, A.P. Linked open social signals. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Toronto, ON, Canada, 31 August–3 September 2010; pp. 224–231.
14. Singh, D.; Tripathi, G.; Jara, A.J. A survey of internet-of-things: Future vision, architecture, challenges and services. In Proceedings of the 2014 IEEE World Forum on Internet of Things (WF-IoT), Seoul, Korea, 6–8 March 2014; pp. 287–292.
15. Jara, A.J.; Olivieri, A.C.; Bocchi, Y.; Jung, M.; Kastner, W.; Skarmeta, A.F. Semantic web of things: An analysis of the application semantics for the iot moving towards the iot convergence. *Int. J. Web Grid Serv.* **2014**, *10*, 244–272. [CrossRef]
16. Lösch, U.; Müller, D. Mapping microblog posts to encyclopedia articles. *Lect. Notes Inform.* **2011**, *192*, 150.
17. Abel, F.; Gao, Q.; Houben, G.-J.; Tao, K. Semantic enrichment of twitter posts for user profile construction on the social web. In Proceedings of the Extended Semantic Web Conference, Crete, Greece, 29 May–2 June 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 375–389.
18. Meij, E.; Weerkamp, W.; de Rijke, M. Adding semantics to microblog posts. In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, Seattle, WA, USA, 8–12 February 2012; ACM: New York, NY, USA, 2012; pp. 563–572.
19. Park, N.; Ko, Y. Computer education's teaching-learning methods using educational programming language based on steam education. In Proceedings of the NPC 2012, Gwangju, Korea, 6–8 September 2012; Volume 7513, pp. 320–327.
20. Oh, J.; Lee, J.; Kim, J. Development and application of steam based education program using scratch: Focus on 6th graders' science in elementary school. In *Multimedia and Ubiquitous Engineering*; Springer: Dordrecht, The Netherlands, 2013; pp. 493–501.
21. Kim, Y.; Park, N. Development and application of steam teaching model based on the Rube Goldberg's invention. In *Computer Science and Its Applications*; Springer: Dordrecht, The Netherlands, 2012; pp. 693–698.
22. Yakman, G. Steam education: An overview of creating a model of integrative education. In Proceedings of the Pupils' Attitudes towards Technology (PATT-19) Conference: Research on Technology, Innovation, Design & Engineering Teaching, Salt Lake City, Utah, USA, 21–23 February 2008.

23. Kim, Y.; Ji, S.; Jung, S.; Moon, J.; Hwang, E. Semantic Enrichment of Twitter News for Differentiated STEAM Education. In Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, 15–18 January 2018; In press.
24. Abbott, J.; President, V. *Nie: Getting Started: A Guide for Newspaper in Education Programs*; Taylor McGaughy Publishing: Reston, VA, USA, 2005.
25. Wu, Y.-C.; Yang, J.-C. Developing a multilingual news reading environment for newspaper reading education. In Proceedings of the 2013 IEEE 13th International Conference on Advanced Learning Technologies, Beijing, China, 15–18 July 2013; pp. 199–203.
26. Choo, H.-P. Online Newspaper in Education (NIE): A New Web-Based NIE Model for Teaching and Learning in School. Master's Thesis, University of Southern California, Los Angeles, CA, USA, 2005.
27. Oliveras, B.; Márquez, C.; Sanmartí, N. The use of newspaper articles as a tool to develop critical thinking in science classes. *Int. J. Sci. Educ.* **2013**, *35*, 885–905. [CrossRef]
28. Wang, Y.-F. Newspapers in science education: A study involving sixth grade students. *J. Educ. Sci. Environ. Health* **2016**, *2*, 98–103.
29. Abbott, J.; Vassilikos, M.; Woodcock, S.; Hendricks, M.; Andrade, S. *Digital NIE. A Guide to Using E-Editions with NIE Programs*; Newspaper Association of America Foundation: Arlington, VA, UISA, 2007.
30. Kim, Y.; Hwang, E.; Rho, S. Twitter news-in-education platform for social, collaborative, and flipped learning. *J. Supercomput.* **2016**, 1–19. [CrossRef]
31. Kim, Y.; Na, B.; Park, J.; Rho, S.; Hwang, E. Twitter News in Education Platform for Collaborative Learning. In Proceedings of the 2016 International Conference on Platform Technology and Service (PlatCon), Jeju, Korea, 15–17 Feberuary 2016; pp. 1–4.
32. Sankaranarayanan, J.; Samet, H.; Teitler, B.E.; Lieberman, M.D.; Sperling, J. Twitterstand: News in tweets. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 4–6 November 2009; ACM: New York, NY, USA, 2009; pp. 42–51.
33. Sriram, B.; Fuhry, D.; Demir, E.; Ferhatosmanoglu, H.; Demirbas, M. Short text classification in twitter to improve information filtering. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, 19–23 July 2010; ACM: New York, NY, USA, 2010; pp. 841–842.
34. Rosa, K.D.; Shah, R.; Lin, B.; Gershman, A.; Frederking, R. Topical clustering of tweets. In Proceedings of the ACM SIGIR: SWSM, Beijing, China, 28 July 2011.
35. Freda, M.C. The readability of American academy of pediatrics patient education brochures. *J. Pediatr. Health Care* **2005**, *19*, 151–156. [CrossRef] [PubMed]
36. Ghose, A.; Ipeirotis, P.G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1498–1512. [CrossRef]
37. White, S. *The 2003 National Assessment of Adult Literacy (NAAL)*; NCES, I.o.E.S., Ed.; U.S. Department of Education: Washington, DC, USA, 2003.
38. Silva, C.; Ribeiro, B. The importance of stop word removal on recall values in text categorization. In Proceedings of the International Joint Conference on Neural Networks, Portland, OR, USA, 20–24 July 2003; pp. 1661–1666.
39. Agrawal, A.; Gupta, U. Extraction based approach for text summarization using k-means clustering. *Int. J. Sci. Res. Publ.* **2014**, *4*, 1–4.
40. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008; Volume 1.
41. Balakrishnan, V.; Lloyd-Yemoh, E. Stemming and lemmatization: A comparison of retrieval performances. *Lect. Notes Softw. Eng.* **2014**, *2*, 262. [CrossRef]
42. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
43. Farrahi, K.; Gatica-Perez, D. What did you do today?: Discovering daily routines from large-scale mobile data. In Proceedings of the 16th ACM International Conference on Multimedia, Vancouver, BC, Canada, 26–31 October 2008; ACM: New York, NY, USA, 2008; pp. 849–852.
44. Coleman, M.; Liau, T.L. A computer readability formula designed for machine scoring. *J. Appl. Psychol.* **1975**, *60*, 283. [CrossRef]
45. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *32*, 221. [CrossRef] [PubMed]

46. Cherry, L.L.; Vesterman, W. *Writing Tools: The Style and Diction Programs*; Bill Laboratories: Manhattan, NY, USA, 1981.
47. Dietze, S.; Sanchez-Alonso, S.; Ebner, H.; Yu, H.Q.; Giordano, D.; Marenzi, I.; Pereira Nunes, B. Interlinking educational resources and the web of data: A survey of challenges and approaches. *Program* **2013**, *47*, 60–91. [CrossRef]
48. Liao, Y.; Lezoche, M.; Panetto, H.; Boudjlida, N. Semantic annotation model definition for systems interoperability. In Proceedings of the on the Move to Meaningful Internet Systems: OTM 2011 Workshops, Crete, Greece, 17–21 October 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 61–70.
49. De Vocht, L.; Softic, S.; Ebner, M.; Mühlburger, H. Semantically driven social data aggregation interfaces for research 2.0. In Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, Graz, Austria, 7–9 September 2011; ACM: New York, NY, USA, 2011; p. 43.
50. Stankovic, M.; Rowe, M.; Laublet, P. Mapping tweets to conference talks: A goldmine for semantics. In Proceedings of the Workshop on Social Data on the Web, Shanghai, China, 8 November 2010.
51. Wikipedia. Ontology (Information Science). Available online: https://en.wikipedia.org/wiki/Ontology_(information_science) (accessed on 18 April 2018).
52. Razis, G.; Anagnostopoulos, I. Semantifying twitter: The influence tracker ontology. In Proceedings of the 2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Corfu, Greece, 6–7 November 2014; pp. 98–103.
53. Oracle. Oracle Spatial and Graph RDF Semantic Graph Developer's Guide. Available online: https://docs.oracle.com/database/121/RDFRM/toc.htm (accessed on 18 April 2018).
54. Tilahun, B.; Kauppinen, T.; Keßler, C.; Fritz, F. Design and development of a linked open data-based health information representation and visualization system: Potentials and preliminary evaluation. *JMIR Med. Inform.* **2014**, *2*, e31. [CrossRef] [PubMed]
55. Hu, Y.; Janowicz, K.; McKenzie, G.; Sengupta, K.; Hitzler, P. A linked-data-driven and semantically-enabled journal portal for scientometrics. In Proceedings of the International Semantic Web Conference, Sydney, Australia, 21–25 October 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 114–129.
56. Togias, K.; Kameas, A. An ontology-based representation of the twitter REST API. In Proceedings of the 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, Athens, Greece, 7–9 November 2012; pp. 998–1003.