Research Article

A Text Analytic Approach to Classifying Document Types

Steven Walczak, *University of South Florida*

Structured Abstract

- **Background**: While it is commonly recognized that almost every work and research discipline utilize their own taxonomy, the language used within a specific discipline may also vary depending on numerous factors, including the desired effect of the information being communicated and the intended audience. Different audiences are reached through publication of information, including research results, in different types of publication outlets such as newspapers, newsletters, magazines, websites, and journals. Prior research has shown that students, both undergraduate and graduate, as well as faculty may have a difficult time locating information in different publication outlet types (e.g., magazines, newspapers, journals). The type of publication may affect the ease of understanding and also the confidence placed in the acquired information. A text analytics tool for classifying the source of research as a newsletter (used as a substitute for newspaper articles), a magazine, or an academic journal article has been developed to assist students, faculty, and researchers in identifying the likely source type of information and classifying their own writings with respect to these possible publication outlet types.

- **Literature Review**: Literature on information literacy is discussed as this forms the motivation for the reported research. Additionally, prior research on using text mining and text analytics is examined to better understand the methodology employed, including a review of the original Scale of Theoretical and Applied Research system, adapted for the current research.

- **Research Questions**: The primary research question is: Can a text mining and text analytics approach accurately determine the most probable

DOI: 10.37514/JWA-J.2017.1.1.06

publication source type with respect to being from a newsletter, magazine, or journal?

- **Methodology**: A text mining and text analytics algorithm, STAR′ (System for Text Analytics-based Ranking), was developed from a previously researched text mining tool, STAR (Scale of Theoretical and Applied Research), that was used to classify the research type of articles between theoretical and applied research. The new text mining method, STAR′, analyzes the language used in manuscripts to determine the type of publication. This method first mines all words from corresponding publication source types to determine a keyword corpus. The corpus is then used in a text analytics process to classify full newsletters, magazine articles, and journal articles with respect to their publication source. All newsletters, magazine articles, and journal articles are from the library and information sciences (LIS) domain.

- **Results**: The STAR′ text analytics method was evaluated as a proof of concept on a specific LIS organizational newsletter, as well as articles from a single LIS magazine and a single LIS journal. STAR′ was able to classify the newsletters, magazine articles, and journal articles with 100% accuracy. Random samples from another similar LIS newsletter and a different LIS journal were also evaluated to examine the robustness of the STAR′ method in the initial proof of concept. Following the positive results of the proof of concept, additional journal, magazine, and newsletter articles were used to evaluate the generalizability of STAR′. The second-round results were very positive for differentiating journals and newsletters from other publication types, but revealed potential issues for distinguishing magazine articles from other types of publications.

- **Discussion**: STAR′ demonstrates that the language used for transferring information within a specific discipline does differ significantly depending on the intended recipients of the research knowledge. Further work is needed to examine language usage specific to magazine articles.

- **Conclusions**: The STAR′ method may be used by students and faculty to identify the likely source of research or discipline-specific information. This may improve trust in the reliability of information due to different levels of rigor applied to different types of publications. Additionally, the STAR′ classifications may be used by students, faculty, or researchers to determine the most appropriate type of outlet and correspondingly the most appropriate type of audience for the reported information in their own manuscripts, thereby improving the chance for successful sharing of

information to appropriate audiences who will deem the information to be reliable, through publication in the most relevant outlet type.

Keywords: information literacy, journal, library information science, magazine, newsletter, STAR′, text analytics, text mining, writing analytics

1.0 Background

While it is commonly recognized that various work and research disciplines have their own language (Luchins, 2007), such as the acronym AI standing for artificial intelligence in the domain of computer science and standing for artificial insemination in the domain of veterinary science[1], the language used within these domains may not be easily recognized by actors from outside of the discipline (Teranes, 2013). Some well-known examples of this discipline-specific language use are: legalese (Kimble, 2013), medicalese (Young, Norman, & Humphreys, 2008), and computerese (McCune, 1999). Additionally, language use within a discipline may vary markedly depending on the intention of the author or speaker, specifically the intended audience and desired perception of trust in the information conveyed.

Within written works, language also differs significantly depending on the type of publication, which is related to the intended audiences and intention of the author(s) for conveying information. Various written forms are commonly used for both research and education purposes. Researchers may elect to publish their findings in journals, books, or magazines and the Internet enables a more uniform access to all forms of scholarly communications (Yancey, 2016). Differences between newspapers, websites, magazines, and journals are highlighted in Table 1. Newspapers, magazines, and journals may be provided as website pages, but the website category is meant to represent informational webpages that are not published in another of the specified formats. An example of an educational use of newspaper articles is an assignment for undergraduate students to find and report on weekly newspaper articles detailing current cybercrime events in an information security course. Newspapers are used to reflect the immediacy of the cyber security events reported.

[1] AI also has numerous other meanings. A few examples are: “as if” in social media/texting; top level domain name for the country Anguilla; and multiple meanings in medicine including adrenal or aortic insufficiency, angiotensin I, and amnioinfusion.

Table 1

Qualities of Various Information Sources

	Newspapers	Websites	Magazine	Journal
Frequency/ periodicity	daily	No periodicity	weekly/monthly	quarterly/annually
Technical level	non-technical	varies	non-technical or small technicality	very technical
Peer review	No	No	sometimes	usually
Audience	general	varies, usually general	general within discipline/topic	specific to sub-discipline
Length	very short, less than 1 page	varies, 1-3 web pages	medium, 1-10 pages	long, 10-30+ pages
Time delay	1 day	immediate	1 week – 6 months	months – year(s)
Examples	*Wall Street Journal*; *USA Today*	Wikipedia.org	*Communications of the ACM*; *Psychology Today*	*Journal of Writing Analytics*; *Journal of Classification*

As shown in Table 1, different types of periodicals target different audiences. Students and researchers utilize newspapers, magazines, and journals for different purposes. The trustworthiness of a source influences the strength of belief and subsequent opinion modification for information gained from a specific source (Hovland & Weiss, 1951). Educators may need to know the source, with respect to the type of outlet, for information being reported in a student research report or for utilization in their own research in order to establish trustworthiness and consequent belief. Similar knowledge will aid in determining the validity and subsequent trust that is needed for utilizing information from an unknown source in developing research hypotheses. Another benefit of having a method for automatically determining the type of publication where information may best be conveyed is to assist student and academic authors in determining appropriate outlets for their research and intended audience (as indicated by the language used in the article).

This article reports exploratory proof of concept research to develop the STAR′ (System for Text Analytics-based Ranking, pronounced star-prime) method, which uses text mining and text analytics to evaluate the source type of textual information and classify the text as an article from either a newsletter, magazine, or journal. The efficacy of the proposed method is evaluated through classification of articles from two newsletters: the Florida Health Science Library

Association newsletter *FHSLAlert* and the *Library Bulletin* newsletter from LSU Health; one magazine: the *D-Lib* Magazine; and three journals: the *Journal of Education for Library and Information Sciences*, the *Journal of Librarianship and Information Science*, and the *Journal of Library Administration*. The STAR′ classification method is able to accurately classify all newsletters taken as a complete unit and all journal articles with 100% accuracy, except for one journal article that is not classifiable when compared to magazines. The initial test of the magazine articles also resulted in 100% accuracy, but further evaluation revealed that overall magazine classification accuracy ranged from 70% to almost 87%.

2.0 Literature Review

According to research by Klosterman, Sadler, and Brown (2012), almost 45% of secondary school educators utilize newspapers, magazines, and other digital information resources in teaching science education, which indicates that college level students are already familiar with these types of information resources for learning. The author offers extra credit to his students in an information security course for bringing in recent newspaper articles detailing cybersecurity events and requires students in a different course to make bi-weekly reports on both magazine articles and websites containing information related to healthcare information technology, which supports the findings of educators' use of various types of publication media in their classrooms. Chu et al. (2016) report that as early as primary school, students identify newspapers as a trusted information source, though they also found that primary school students' first stop when information seeking was a general web search engine. University research projects frequently require students to use either specific or multiple types of information sources, such as a book and multiple journal articles (Jamieson, 2016).

University students are relying ever more heavily on the World Wide Web (Web) as an information source for completing course projects, including electronic versions of scholarly and popular media, but research has shown that modern students rely on this information without verification and may be using biased and inaccurate information from Web sources (Barclay, 2017; Metzger, Flanagin, & Zwarun, 2003). Kissel et al. (2016) report that university students' primary research sources are the Internet, web search engines, or magazines, with less than 20% of the students indicating they could evaluate the credibility of a source. Librarians typically recommend books, journal articles, government documents, and specialized news sources as appropriate scholarly sources for undergraduate research papers (Jamieson, 2016).

Appropriate knowledge of information resources and the ability to evaluate the credibility of these resources is a necessary skill for students in both

their academic journey and future careers. Employers expect job applicants, including recent graduates, to be able to utilize information in a variety of formats and develop a deep understanding of source and related information in addition to being able to quickly acquire information from electronic sources (Head & Wihbey, 2014).

2.1 Information Literacy

Information literacy is a critical skill for university student success (Laubersheimer, Ryan, & Champaign, 2016). How are students accessing information and evaluating information from various resources to meet their educational and future professional informational literacy needs?

A search engine such as Google is the first stop for the majority of students searching for project or research information, and few use academic resources, specifically journal article catalogues, for locating information (Griffiths & Brophy, 2005; Kissel et al., 2016). This means that students are relying on the ranking of information resources by search engines to determine the quality of information retrieved. Academic faculty also fall into this behavioral trap by frequently ignoring journal articles and other relevant sources of information if they are not readily available electronically (Bonthron et al., 2003). In general, 90% of search engine users will only click on a link from the first page of results, which typically contains only 10 results (Wiebe, 2016).

A risk associated with treating all types of publications as equally reliable and relevant is that conference papers, newsletter articles, magazine articles, and invited journal articles do not go through the same review process as traditional academic journal articles (Franco, Malhotra, & Simonovits, 2014) and as such may contain unreproducible or unverifiable results. A study by Reed (1999) showed that in three occupational therapy journals, over 61% of the citations were to other journal articles, but that newspaper articles, newsletter articles, and conference proceedings, among other non-scholarly resources, made up just under 13% of all citations. The Association of Research Libraries maintains a directory of electronic journals and newsletters, but elected in 1995 to focus on more scholarly publications and as a result, dropped the cataloguing of newsletter articles and conference proceedings from that date forward (Mogge, 1999), thus clearly defining newsletters as being less reliable information resources for scholarly investigation and research. Numerous prior research studies identify the inability of current students, especially at the undergraduate level, to evaluate the validity and reliability of information obtained electronically (Barclay, 2017; Davis & Cohen, 2001; Wiebe, 2016). Additional research states that students do not understand the difference between journal articles and less reliable sources

like magazines and web pages and are often unable to accurately identify the source type (Laskin & Haller, 2016).

Differences in perceptions of the reliability of information is not a new problem engendered by the rapidly growing content available via the Internet. Worsley (1989) examined how health consumers rated newspapers, magazines, and commercial information as being much less reliable than traditional medical resources for gathering health information. Other research has shown that after less than a decade of publicly available Internet browsers and search engines, undergraduate students' citations of websites and newspaper articles have increased, while their concurrent citations of magazine and journal articles have decreased (Davis & Cohen, 2001). Traditional print media and its electronic counterparts, as shown in Table 1, provide different levels of quality and reliability of information due in part to the presence or lack of peer review, as well as the intended audience of these different publication outlets.

Scientific information may be obtained from different formats, with high school students relying more on textbooks and adults relying more on magazines and newspapers (Tseng et al., 2010). Scientifically literate adults should be able to understand reports from various media and be able to discuss the merits of the findings.

Cockrell and Jayne (2002) conducted an experiment in which undergraduate students, graduate students, and faculty were asked to find information that appeared in one of three possible publication outlets: newspapers, magazines, or journals. All three subject types were most successful at finding information appearing in newspapers. Graduate students and faculty, although less proficient with magazines and journals than newspapers, did not show any difference in their proficiency of finding information contained in these two types of publications. However, undergraduate students were much more successful at finding information when it appeared in a magazine versus a journal source. Students frequently have difficulty citing information that they find in electronic resources (Davis & Cohen, 2001). These results indicate that not only do people have difficulty determining the reliability of information, but they also have difficulty distinguishing between these three different types of information sources.

Due to the evolving dynamic of information availability enabled through technology (Klosterman, Sadler, & Brown, 2012), Barclay (2017) claims that information sources should not simply be classified as good or bad. The authority and credibility of information resources should be viewed contextually and is dependent on how the information is used and the desired outcomes and audience of such use (Lloyd, 2005). Thus, the inherent need to develop a classification methodology to determine the likely source of information with regard to

newspaper, magazine, or journal articles in order to evaluate the efficacy of such information in being perceived as credible and reliable. Furthermore, identification of the most appropriate type of publication outlet will enable greater absorption and knowledge acquisition from the information presented by targeting relevant and receptive audiences (Kellogg & Walczak, 2007; Susman & Evered, 1978).

2.2 Text Mining and Text Analytics

A heuristic approach to information source classification that utilizes text mining and text analytics is proposed in this article. Hence, a brief review of text mining and text analytics is presented.

Text mining and text analytics is an important technique with growing interest from researchers (Sebastiani, 2002). Analysis of research literature is increasingly being performed with the usage of text mining and text analytics methods (Bragge, Thavikulwat, and Töyli, 2010). Text analytics has also been identified as one of the primary methods applicable to performing analytics on large sets of unstructured data (Gandomi & Haider, 2015), such as that found in collections of newspaper, magazine, and journal articles. Tseng et al. (2010) utilize text analytics techniques to identify scientific concepts from news stories. Text mining techniques may be used to identify specific properties or elements of a document (Tseng et al., 2010) or more commonly to classify the entire document (Bichindaritz & Akkineni, 2006).

The Internet has made enormous quantities of data available, and this creates a need for new methods to dynamically discover knowledge based on interesting patterns found in unstructured text data (Aggarwal & Zhai, 2012). Text mining seeks to extract information from text sources that potentially have a very large number of words which may display vagueness, uncertainty, and fuzziness in relation to their interpretation through more programmatic means (Hotho, Nürnberger, & Paaß, 2005).

Text mining first requires tokenization, which removes all punctuation and other non-text characters and replaces each character or a consecutive sequence of non-text characters with a single delimiting character, typically a space (Hotho, Nürnberger, & Paaß, 2005). Merging all the tokenized words from a collection of documents creates a dictionary for that document set. The next step is to identify important/meaningful words within the dictionary to perform further text analytics. Text analytics commonly identifies words that are meaningful to a particular domain and hence contained within the corpus, through use of term frequency (*tf*) and document frequency (*df*) across a collection of documents (D), where term frequency is simply the average number of times a term from the dictionary occurs within the document set and document frequency is the number

of documents within the set that contain at least one occurrence of a tokenized word (Yamamoto & Church, 2001).

A commonly used text mining approach is the bag-of-words approach in which a collection of words is used to define a corpus from words with both high term frequency and high document frequency, and the words in that corpus are then used in classifying documents (Crain et al., 2012). Document classification may use term specificity or inverse document frequency (IDF) as defined in equation (1) as a part of the TF*IDF (defined in equation (2)) method for classifying new documents with respect to the corpus (Robertson, 2004). The bag-of-words assumption implies that the order of the words found within a document is not important, merely their presence is significant, which is consistent with how informative articles are typically written (Goutte, 2008).

$$\text{IDF} = \frac{-log_2\, df}{\text{D}} \quad (1)$$

$$\text{TF*IDF} = tf * \text{IDF} = \frac{tf * -log_2\, df}{\text{D}} \quad (2)$$

Other types of text mining document analysis may use other probabilistic and logical models (Robertson, 2004; Sun, Deng, & Han, 2012), Bayesian reasoning (Kim et al., 2006; Sun, Deng, & Han, 2012), or other machine learning based models such as artificial neural networks (Isa, Kallimani, & Lee, 2009) to derive meaning. Text analytics has been shown to outperform decision tree methodologies in classification of documents (Fan et al., 2006). The STAR′ text mining and analytics method uses *tf*, *df*, logical, probabilistic, and heuristic reasoning.

2.3 Original STAR algorithm

The STAR′ text mining and analytics method is based on Walczak and Kellogg's (2015) previously published STAR (Scale of Theoretical and Applied Research) algorithm. The original STAR algorithm was used to collect and analyze regular expressions from journals in the business analytics domain. These regular expressions were then used to identify terms from journals that published primarily theoretical research and journals that published primarily applied research. Their results (Walczak & Kellogg, 2015) indicated a very high correlation with the evaluations from a panel of experts, the editors of the 23 selected business analytics journals, achieving a mean error of 0.696 on a scale of one to seven.

A regular expression is a way to define patterns of text as being considered identical and is a commonly used technique within text mining to help reduce the total number of words contained within a dictionary (Dietrich, Heller, & Yang, 2015). For example, the regular expression calculat* would match all of

the words: calculate, calculated, calculates, calculating, calculation, calculations, calculator, and calculators. The regular expressions contained in STAR′ use wild cards that may be used to match zero or more of any character, one of any character, or to force a regular expression pattern to appear at the beginning or end of a word. Care must be taken when using the wild card in a regular expression to ensure that it does not over-generalize and capture terms as equivalent, which in fact have significantly different meanings (e.g., famil* captures family and families as well as familiarize, familiarities, familiarity, etc., which have different meanings). Regular expressions have other capabilities not utilized in STAR′ and the interested reader is directed to Fitzgerald (2012).

Text scraping and tokenization typically aim to collect all of the words present in a document. Research has shown that the most commonly occurring words are generally not meaningful in determining the document, genre, author or other document-relevant information for a specific document or set of documents as these terms occur in all documents, such as the twenty most common words in Shakespeare's *Hamlet* (Dietrich, Heller, & Yang, 2015), shown in Table 2. As seen in Table 2, the word "the" was the most common word in Hamlet, and "the" is also likely to appear in most text documents analyzed. The IDF technique mentioned above claims that the less frequently a term appears, the more meaningful it will be (Yamamoto & Church, 2001), and while this is not precisely how the STAR′ algorithm treats its collected words, from regular expressions, it does eliminate non-meaningful words including: articles, pronouns, prepositions, conjunctions, exclamations, adjectives, adverbs, numeric words, and any word less than three characters in length (e.g., is, be).

Table 2

Twenty Most Common Words in Shakespeare's Hamlet *(Dietrich, Heller, & Yang, 2015, p. 266)*

the	and	of	to	my
a	in	I	you	is
his	it	not	your	that
with	this	be	he	for

The original STAR text mining and text analytics methodology is separated into three distinct phases (Walczak & Kellogg, 2015), and this approach is mimicked in the new STAR′ method. First a collection of documents is identified to develop the classification dictionary. This collection must have representative documents from each category to be classified. Traditional text mining then proceeds in phase one by advancing through each text document one

character at a time and tokenizing the text. Character by character processing of text documents to perform tokenization is a popular method of identifying word (and also regular expression) boundaries (Yamamoto & Church, 2001). Once all words, except those that are automatically eliminated as non-meaningful, are identified, they are stored into an Excel™ spreadsheet.

After all documents have their dictionary words extracted and placed into spreadsheets, phase two begins and documents of the same category have their respective dictionary words and occurrence counts, *tf*, alphabetized and then placed into adjacent columns in a single spreadsheet. Phase two currently involves manual examination of the terms to select those terms that will advance into the corpus for that document category. In order for a regular expression, hereafter called a word pattern, to be placed into one of the two corpuses, it must satisfy the following four heuristics:

1. The word pattern must occur at least α times on average across all documents, $\frac{tf}{\mathrm{D}}$, for that category;
2. The word pattern must appear in at least *N* documents for that category, *df* ;
3. The word pattern must appear in at least β different journals; and
4. The word pattern may not appear in more than γ documents of the other category.

The original classification of journal articles into theoretically-oriented versus application-oriented research types utilized 15 randomly selected journal articles from each research classification, three distinct journals for the theoretical and two distinct journals for the applied, and used the following values: α = 0.667, *N* = 3, β = 3, γ = 1. These values were pre-set in the algorithm. A higher α value would correspond to lower IDF, but also indicates that to be used as a classifier, a potential word must have a high enough *tf* value that would assist in eliminating chance occurrences of word patterns. An *N* of 3 indicates that the IDF value derived from equation (1) had to have a minimum value of 0.405. The β value is used to make the identified word patterns more generalizable and prevent patterns that are localized to a single journal. Lastly, the γ value indicates that identified terms are not required to be unique to their respective categories, as this would be achieved with a γ value of zero, but should only appear very infrequently in documents from the other category.

Phase three of the STAR text analytic methods uses the identified corpuses to determine the relative similarity of a new document with the specified categories (Walczak & Kellogg, 2015). As may be seen from the heuristics above, the STAR algorithm is used to classify documents into one of two possible

categories. However, the STAR algorithm produces a continuous value as opposed to a strict binary classifier to show the relative representativeness of the newly classified document with the specified category, thus creating a closeness criterion similar to nearest neighbor approaches to classifying text documents (Baoli, Qin, & Shiwen, 2004). Values ranged from -200 for pure theoretical research to 200 for pure applied research, with a value of zero representing an equally balanced approach. The larger the absolute value of the STAR prediction value implies greater certainty in the correctness and uniqueness of the specified classification.

Assume that a single regular expression contained in the corpus R_i is denoted by $r_{i,j}$, where j denotes the j^{th} term within the i^{th} corpus, hence $r_{i,j} \in R_i$. The subscript on the corpus *R* denotes that the STAR text analytics method requires two corpuses, so i has a value of 1 or 2, where the first corpus will generate negative values and the second corpus will generate positive values. Any number of corpuses may exist, each produced by analyzing a different class of documents during phases one and two of the text mining aspect of STAR, but only two at a time of these corpuses are used for performing the classification. Assume that S_i represents the size or number of elements contained within corpus i. The presence of a regular expression in a document is denoted by a document frequency, $df(r_{i,j})$, being set to the value 1, otherwise 0 if not present in the document. Each regular expression also has a term frequency for the analyzed document, denoted $tf(r_j)$, which represents the number of times that regular expression occurs within the analyzed document. The calculation of the STAR classification value is given in equation (3).

$$\text{STAR} = a\left(\sum_{j=1}^{S_2} df(r_{2,j}) - \sum_{j=1}^{S_1} df(r_{1,j})\right) + b\left(100\frac{\sum_{j=1}^{S_2} df(r_{2,j})}{S_2} - 100\frac{\sum_{j=1}^{S_1} df(r_{1,j})}{S_1}\right) + c\left(\frac{\sum_{j=1}^{S_2} tf(r_{2,j})}{\sum_{j=1}^{S_2} df(r_{2,j})} - \frac{\sum_{j=1}^{S_1} tf(r_{1,j})}{\sum_{j=1}^{S_2} df(r_{1,j})}\right) \quad (3)$$

Equation (3) uses three constants, *a*, *b*, and *c* which may be adjusted to account for different term relevance within different document classifications. The first part of equation (3) accounts for the frequency of the presence of the various terms from each corpus within a document. The second term adjusts for potential discrepancies when comparing *df* from two unequal sized corpuses where one is significantly larger than the other. Research has shown that the mere presence of terms may be insufficient for accurate classification of documents and that *tf* may be more reliable (Conway, 2010). Hence, the third term of the equation introduces relative term frequencies across both corpuses for the document being analyzed. A plain description of the equation would be: the

difference in term presence, plus the difference in coverage from the corpuses, plus the difference in relative term frequency within the document.

3.0 Research Question and Proposed Evaluation Method

As noted above, information literacy is strongly affected by the perceived reliability of the information source, and significant differences exist in the perceptions of reliability between newspaper articles, magazine articles, and journal articles. Text mining and text analytics offer an opportunity to automatically classify information sources with regard to type of publication and consequent perceived reliability of the information between newspapers, magazines, and journals. Text mining and text analytics approaches have already shown efficacy in identifying research articles from the PubMed database to assist in assessing chemical health risks (Korhonen et al., 2012), finding relevant scientific information from newspaper articles (Tseng et al., 2010), identifying scientific classification for organizing research proposals (Ma et al., 2012), and identifying the theoretical or applied nature of research reported in journal articles (Walczak & Kellogg, 2015).

With regard to information literacy, there is a need to be able to quickly and accurately distinguish the source of information presented in both student and faculty research. While prior text mining research has shown the ability to identify relevant articles from newspapers, the broad nature of newspaper coverage makes it problematic when using a text analytics corpus developed for a particular domain. Therefore, newsletters were substituted for newspapers in this research because, while they have similar types of coverage as newspapers (Yao, 2009), they are more focused within a specific domain and will consequently utilize language more consistently across articles and issues of the newsletter.

The primary research question, which is derived from these issues, is: can a text mining and text analytic method accurately identify the source of a publication with respect to journal articles, magazine articles, and newspaper/newsletter articles? The purpose of this research question is to find an accurate method for distinguishing between more reliable and less reliable information resources and also to be able to identify appropriate sources, when a specific source type is specified as required in an assignment for students.

The new STAR′ text mining and analytics methodology was developed from Walczak and Kellogg's (2015) original STAR classifier. The current version of STAR′ is a two group neighborhood classifier and is used to classify articles from the domain of Library and Information Sciences (LIS) into their relevant publication type: journal, magazine, or newsletter. As such, this produces three hypotheses from the research question above:

H_1 : The STAR′ text analytics method accurately distinguishes between journal articles and newsletter articles (in the LIS domain);
H_2 : The STAR′ text analytics method accurately distinguishes between journal articles and magazine articles (in the LIS domain); and
H_3 : The STAR′ text analytics method accurately distinguishes between magazine articles and newsletter articles (in the LIS domain).

The LIS domain is specified in the research hypotheses in parentheses because this is the actual domain of the journal, magazine, and newsletter articles examined in the current research. Language usage differences occur across research and teaching domains, but also within domains across different contexts, where the publication outlet or intended readership serves as a type of context. "Information literacy requires engagement with information through the discourse and discursive practices specific to the context (Lloyd, 2005, p. 84)." The same hypotheses should be generalizable to other domains and thus allow removal of the hypotheses portions specified in the parentheses, with the prerequisite qualification that documents from the corresponding domain are analyzed in phase one of the STAR′ text mining method.

Distinguishing between each of the two means that the STAR′ method will correctly classify examples of each type of information publication source. Prior research has stated that the optimal way to evaluate text mining classifiers is through prediction accuracy, which is determined by the number or percentage of correct predictions made by the classifier (Cohen & Hersh, 2005). Classification accuracy will be the average correct classification accuracy for each of the two publication outlet types evaluated in each of the three combinations of different classifications: journal or newsletter, journal or magazine, and magazine or newsletter.

4.0 Research Methodology

Before text mining can occur, a target collection of documents must be identified. Developing a text mining and text analytics method for classifying whether an article comes from a journal, magazine, or newsletter requires sample documents of each type. One journal, one magazine, and one newsletter, all from the LIS field, were chosen, and articles were then randomly selected from issues published from January 2015 through December 2015 for corpus creation and between January 2016 and December 2016 for classification. The journal is the *Journal for Education in Library and Information Science* (*JELIS*), the magazine is *D-Lib Magazine* (*D-Lib*), the magazine of Digital Library Research, and the newsletter is the newsletter of the Florida Health Science Library Association, *FHSLAlert Newsletter* (*FHSLA*).

The *FHSLA* newsletter is only published twice a year and the articles are very small. For example, for both 2015 issues, a total of 28 articles were published at an average word count of 297.5 words per article, ranging from an article of 83 words up to an article with 896 words. For this reason, the *FHSLA* newsletter was not separated into articles, but rather each issue was treated as a single document to ensure that sufficiently large dictionaries were available for creation of the newsletter corpus. In order to ensure a sufficient quantity of documents for validation of the text analytics method, newsletters were also collected from 2014. One of these 2014 newsletters was used for the text-mining dictionary creation corpus development set and the other was used for the classification validation set, creating two sets of three independent documents from this newsletter source. Because these represent all newsletter documents from the time period, this should be considered a purposive convenience sample for the newsletters only, which is appropriate for exploratory research (Teddlie & Yu, 2007).

Because the analysis of the text analytics method is calculated based on the correct classifications of each set of documents, so as not to bias the results, the corpus development and classification validation sets for the journal and magazine articles are kept to the same size. The three-year time span of article and newsletter selection was maintained to counter extraneous effects from re-framing (e.g., through changes in editor or publisher) of the journal, magazine, or newsletter (Yao, 2009), in which the re-framing might change the language used in the specific outlet.

The collection of documents for each publication outlet type is downloaded in PDF format. These PDF documents are then converted to text documents to remove special codes placed into documents by various word processors for formatting and other metadata, which is a common practice in text analytics (Bui, Del Fiol, & Jonnalagadda, 2016). Journals have elements that are not consistently found in other types of documents, such as abstracts and references. While references may exist to a smaller extent in magazines than journals, to create an unbiased comparison of the classification capabilities of the proposed STAR′ text analytics method, abstracts and reference lists are removed when present from the documents. This should not affect the classification accuracy, since abstracts present redundant information to the reader of what is contained in the body of the article.

The STAR′ heuristic text mining and text analytics method is based on the original STAR algorithm of Walczak and Kellogg (2015) and performs the phase one text mining in the same way. This includes automatic disposal of certain classes of regular expression word patterns: pronouns, predicates, articles, conjunctions, adjectives, adverbs, numbers, numeric words, and any regular

expression pattern shorter than three characters. The collected dictionary words are then sorted by *tf* so that if selected for the corpus, they will be stored in descending order of *tf* magnitude. The text analytics part of STAR′ which will be applied in phase three, must analyze a document's text on a character by character basis to determine if it matches any of the stored regular expression word patterns. The descending *tf* magnitude ordering of the regular expressions lets STAR′ first compare possible word patterns against the most frequently occurring patterns for that outlet type, thus increasing the efficiency of the method.

The phase two corpus creation from the dictionary of words for each set of text mined documents uses the following corpus determination values: $\alpha = 1.667$, $N = 2$, $\beta = 0$, $\gamma = 0$. This α value is larger than the value originally proposed by Walczak and Kellogg (2015) due to the small number of documents being used for corpus development, and consequently requires that the regular expression word patterns appear in documents more frequently to increase the likelihood that they will be repeated in similar outlet type (i.e., journal, magazine, or newsletter) documents. An *N* value greater than 1 is required to make sure that a word pattern is actually repeated across documents of the same outlet type. Because only three documents are currently being text mined by STAR′ to capture dictionaries for the creation of the corresponding corpus, *N* may have a value of either 2 or 3, but a value of 3 would require that corpus selected word patterns appeared in every document that has undergone text mining. However, this is too restrictive and would imply that authors of similar outlet type documents must all use these specific words. All documents for a specific outlet type are taken from the same publication, though from different issues of that publication, which makes the use of any β value greater than 0 meaningless. Future research that examines developing corpuses for outlet type classification which use multiple different journals, magazines, and newsletters could then implement a non-zero β. The γ value of 0 used by STAR′ for the current outlet type classification problem requires that terms identified for each corpus must be unique across the three document samples text mined to create the corpus. This is a reasonable requirement given the small number of documents used, which was predicated by the low quantity of available documents for the newsletter selected for this exploratory study over the specified time period. This does not mean that the corpus words cannot appear in documents of the other types that are outside of the sample used to create each corpus, just that for purposes of corpus creation the terms are required to be unique.

Once the three corpuses are created, the phase three classification of out-of-sample documents is performed. After converting a document to lower case, STAR′ performs a character by character text mining search of the entire document to identify all possible regular expression word patterns contained in

either of the two corpuses currently under consideration for the classification decision. The STAR′ heuristic classification of document publication outlet type utilizes inductive inference (Goutte, 2008), based on text content of documents and similarity to other documents based on occurrence of the regular expressions. A similar calculation as shown in equation (3) is performed to classify documents utilizing the variations in *tf* and *df* across different document outlet types. The *a*, *b*, and *c* constants are modified to reflect values consistent with the LIS domain and typical word utilization expectations within LIS research documents.

Every combination of journal, magazine, and newsletter is evaluated and classified using the STAR′ text analytics method. This means that STAR′ must be run twice for each document to determine its corresponding classification to each of the two other classes of document outlet types. Because the relative *tf* and *df* values are different for each publication outlet type, it may be inferred that the STAR′ classification value will vary for a specific document depending on the other class to which it is being compared, but should still place the document in the correct classification. Results for each classification are then evaluated for accuracy and compared against each other to determine overall classification accuracy.

5.0 Results

The number of unique words text mined from each of the documents to go into the document dictionaries during phase one is shown in Table 3. As would be expected, magazine articles tend to have fewer words than journal articles. The newsletter word counts are on average similar to the journal word counts, but recall that this is for the entire newsletter due to the extremely small article size.

Table 3

Unique Word Patterns Retrieved by STAR′Text Mining

Document Outlet Type	Document 1	Document 2	Document 3	Average word count
Journal (*JELIS*)	830	964	545	780
Magazine (*D-Lib*)	382	363	979	575
Newsletter (*FHSLAlert*)	1051	305	978	778

The mined word patterns were saved into an Excel™ spreadsheet where the phase two analysis of the dictionary words was performed to determine if any

satisfy the heuristic constraints previously listed to become a word in the corpus for that outlet type. The corpuses for LIS domain journals, newsletters, and magazines are shown in Appendix A. Regular expressions within each corpus are represented in lowercase characters because, as noted previously, all text was converted to lowercase prior to matching document text against the corpus regular expression word patterns. The size of each corpus is expected to be relatively small, due to the limited amount of data used to develop each corpus in this exploratory study, but it is important to realize that the subsequent text analytic classification of the document outlet type is being accomplished using only 21 to 26 identifiers. Prior research has shown that a small corpus may still yield a robust classifier (Iriondo et al., 2009; Walczak & Kellogg, 2015; Weiss et al., 1999).

A clarification of the operation of some of the corpus creation heuristics: The acronym "erm" appeared in a single *JELIS* article 89 times, thus satisfying the required α value. However, the acronym pattern "_erm_" (where the underscores indicate a white space character required) did not appear in either of the other two *JELIS* article dictionaries, and as such did not satisfy the *N* heuristic requirement and thus will not be considered as a corpus regular expression pattern for LIS journal articles. The regular expression word pattern librar* occurs in 8 of the 9 documents used for text mining the dictionaries that are used to evaluate the word patterns for inclusion in an individual document type corpus, with individual document *tf* values shown in Table 4. Since this librar* word pattern is not unique to a single outlet type, it fails the γ value for the fourth corpus development heuristic and is not included in any corpus. This makes sense for the domain of LIS where the terms library, librarian, and librarianship are commonly used and cannot therefore distinguish between different publication outlet types.

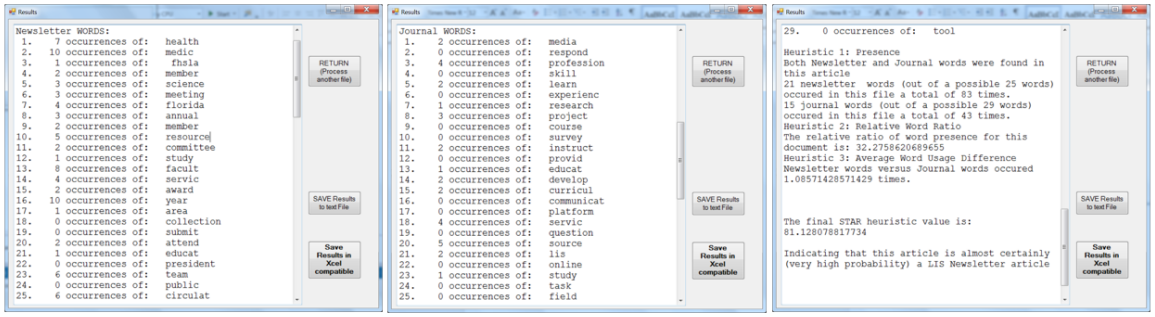
Table 4

Occurrence (tf) of Regular Expression librar in Corpus Training Documents*

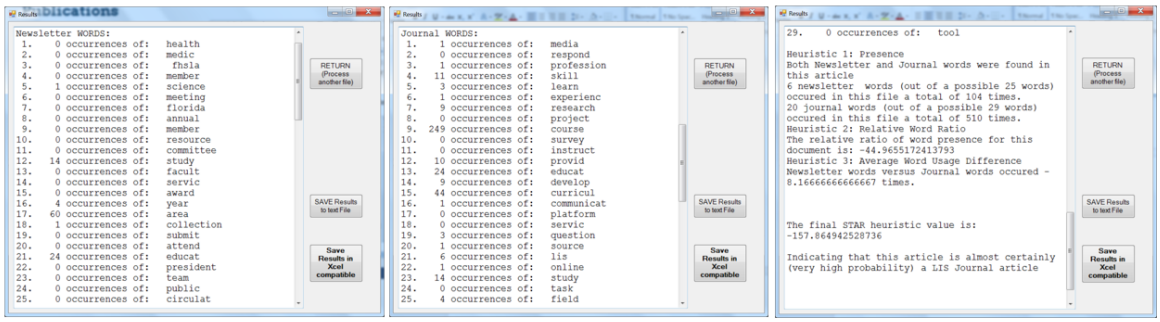
Document Outlet Type	Document 1	Document 2	Document 3
Journal (*JELIS*)	14	104	2
Magazine (*D-Lib*)	1	0	57
Newsletter (*FHSLAlert*)	14	45	105

From Appendix A, the pattern "fhsla" appears in the newsletter corpus. This is due to the presence of the organization name appearing in every newsletter, frequently multiple times. This pattern will almost certainly not be present in any other newsletter from different libraries or other LIS organizations, but is maintained in the corpus because it satisfied all the heuristic criteria used for corpus selecting patterns from the document dictionaries for populating the corresponding corpus. The lack of presence of this word in other newsletter publication documents will diminish its respective similarity score; however, the other newsletter corpus words should suffice to distinguish a newsletter document from other types of publication outlet types.

A sample of the output for the STAR′ classifications phase three tool, written in C#, is shown in Figure 1 for classifying if a document is a newsletter or a journal article. The output screens would be similar for both other classification pairings, journal or magazine and newsletter or magazine, but with different *tf* and calculated heuristics that make up the various parts of the classification value. They would also have different explanations for the similarity-based classification value.



(a) *Florida Health Science Library Association* newsletter issue



(b) *Journal of Education for Library and Information Science* article

Figure 1. Sample screens of STAR′ for a (a) newsletter article and (b) journal article.

The results of comparing each of the three out-of-sample validation cases for the journal and newsletter classification are shown in Figure 2, with the results

for the journal and magazine classification shown in Figure 3, and the results of the magazine and newsletter classification shown in Figure 4. As seen in Figures 2-4, the classification accuracy of STAR′ was 100% for this small proof of concept validation data, with STAR′ classifying each document correctly in both sets of comparisons.

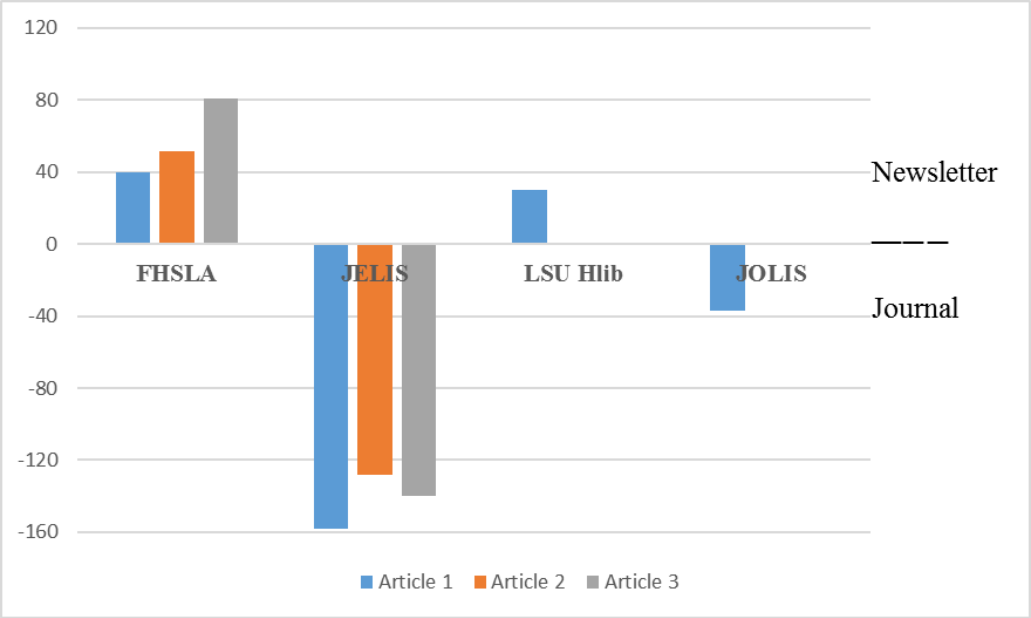


Figure 2. STAR′ classification results for journal articles and newsletters.

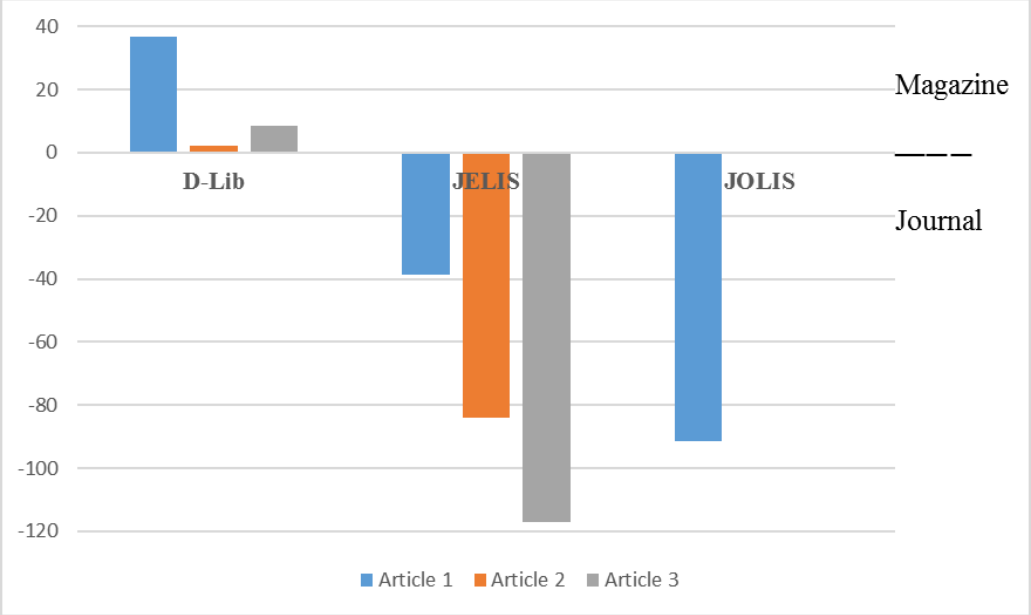


Figure 3. STAR′ classification results for journal articles and magazine articles.

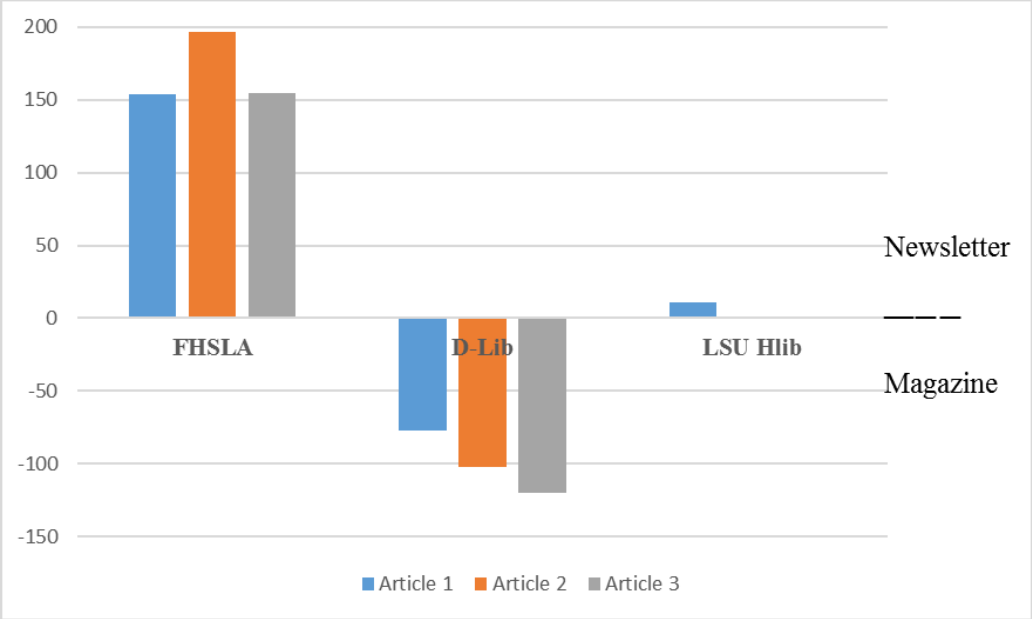


Figure 4. STAR′ classification results for journal magazine articles and newsletters.

A test of the generalization of the STAR′ methodology to other journals and newsletters was performed during the stage one proof of concept analysis by randomly selecting a 2016 article from a different journal and a 2014 newsletter. The journal chosen was the *Journal of Librarianship and Information Science* (*JOLIS*) and the newsletter chosen was the *LSU Health Library Newsletter* (*LSU Hlib*). A 2014 publication date was used for the newsletter because this is the most recent issue available publicly from the website. The values for each of these additional documents are also shown in the respective Figures 2-4. As indicated above, the similarity-based classification differs depending on which other publication outlet type is competing for a document's classification. Both of these additional documents were also classified correctly with 100% accuracy.

Based on the initial results, as shown in Figures 2-4, a second stage validation was performed. For the second stage validation, all remaining 2016 articles from both the *JELIS* and *JOLIS* journals were collected and evaluated, producing an additional 28 journal articles. A new 2016 *FHSLA* newsletter published following the initial test, *FHSLA* newsletters from 2013, and the remaining 2013 and 2104 *LSU HLib* newsletters were collected and evaluated, producing an additional 7 newsletters. All remaining 2016 *D-Lib* articles were also collected and evaluated, producing an additional 27 magazine articles. A new test was performed for the *Journal of Library Administration* (*JLibAdmin*), which although called a journal, is published eight times a year, thus better matching the publication frequency of a magazine. The STAR′ heuristic text classification method was used to help disambiguate if *JLibAdmin* should be considered a

proper journal with a higher than expected publication frequency or a magazine. All articles from the last four issues of *JLibAdmin* from 2016 were collected and analyzed, producing 19 articles for classification.

Results of the second stage analysis are displayed in Table 5, with the *JLibAdmin* articles considered as journal articles. Table 5 clearly shows that for the journal classifications and the newsletter classifications a 0% misclassification rate and a near 100% classification accuracy were achieved, with a single *JOLIS* article unable to be classified. A result of unclassified or not classifiable occurs when the STAR′ heuristic value is between -1.0 and 1.0, indicating that not enough differentiation existed between the regular expression word patterns to enable STAR′ to accurately distinguish between either classification. Users may adjust the heuristic cutoff value for insufficient evidence to classify to be greater than 1, which in turn would require stronger and perhaps more trustworthy evidence of an article's type. For example, adjusting the unknown cutoff to an absolute value of 5.0 would not have affected any of the journal or newsletter article classifications, but would have changed two correct magazine classifications versus journals and one incorrect magazine classification versus newsletters to unknown.

Table 5

Second Stage Evaluation Results of STAR′Classification.

The absolute average STAR′ value is shown below each number of predictions in [ ].

	Journals (N = 47)	Magazine (N = 27)	Newsletters (N = 7)
Correct vs. Journal	N/A	18 (66.67%) [33.67]	7 (100%) [54.5]
Incorrect vs. Journal	N/A	8 (29.6%) [18.03]	0
Unknown vs. Journal	N/A	1 (3.7%) [0.35]	0
Correct vs. Magazine	46 (97.87%) [83.03]	N/A	7 (100%) [134.9]
Incorrect vs. Magazine	0	N/A	0
Unknown vs. Magazine	1 (2.13%) [0.78]	N/A	0
Correct vs. Newsletter	47 (100%) [95.04]	23 (85.19%) [72.44]	N/A
Incorrect vs. Newsletter	0	4 (14.81%) [14.32]	N/A
Unknown vs. Newsletter	0	0	N/A

Table 5 also clearly demonstrates that while the publication frequency of *JLibAdmin* is greater than normally expected for journals, the content of this publication clearly aligns well with a journal. The language used to convey information in *JLibAdmin* is similar to the language used in other LIS journals for conveyance of scientific LIS research.

The magazine classification accuracy was much better than randomly guessing (70% versus journals and 86.67% versus newsletters when combined with the original proof of concept results), which would result in an average accuracy over the long run of 50%. The average of the STAR′ prediction values shown in Table 5 for magazine article predictions indicates that the values are greater, showing stronger evidence, for correct predictions over incorrect predictions. This indicates that while incorrect predictions were made by STAR′ for magazine articles, since the values were typically lower than those achieved for the accurate predictions, further research examining the optimization of the heuristic *a*, *b*, and *c* constants used in the prediction algorithm (see equation (3)) might improve the prediction accuracy for magazine articles.

6.0 Discussion

The classification accuracy for the journal and newsletter outside of those used to develop the corpus implies that the method is generalizable to a wide variety of journals, magazines, and newsletters. The *LSU Hlib* newsletters did not contain the corpus word fhsla, which is specific to the newsletter used to mine the regular expression patterns contained in the corpus, and thus would have reduced coverage of the corpus, which is one of the heuristic factors used to determine the classification. However, the language used in the *LSU Hlib* documents is sufficiently similar to the language of the *FHSLAlert* newsletter and corresponding corpus words to allow for correct classification. Similarly, the *JOLIS* and *JLibAdmin* articles were correctly classified by the regular expression patterns in the journal corpus that were derived from a different journal, indicating the similarity of language, specifically word choice and usage, for this type of highly reviewed and reliable outlet type. Future research will need to examine the development of outlet source corpuses utilizing multiple different examples of appropriate documents from a variety of publishers. It is expected that this future research will continue to show very high classification accuracy and will be even more generalizable for classifying other documents from outlets not participating in the corpus creation for that outlet type.

These results lend evidentiary support for each of the three hypotheses, H_1, H_2, and H_3, confirming the potential for text analytics to identify publication outlet source type. The STAR′ text mining and text analytic similarity-based classification method is able to correctly identify the source of an article as

coming from a journal or newsletter, at least for the data examined in this exploratory study. Additionally, the STAR′ classification method is able to accurately classify a majority of magazine articles.

The *D-Lib* article classifications necessitate further consideration. The sign of the classification of the second *D*-Lib article shown in Figure 3 does accurately place this document in the magazine class; however, the absolute value of the similarity-based classification is just over a value of 2, which receives a text explanation of: "Indicating that this article stands a slightly better chance than pure balance of being a LIS magazine article." If the unclassifiable cutoff value was raised to 3, then this article would have been unclassified, since absolute values less than the cutoff value (currently set to 1) are regarded as not being able to distinguish adequately between the two different publication outlet types. In other words, though the D-Lib article is considered a magazine article, it is not far from the quality and perceived reliability of a journal article, based on the words used to convey the information presented in the article. Why are only two-thirds of the *D*-Lib articles correctly classified when compared against the journal corpus? A possible explanation is that *D-Lib* is actually one of those magazines that is respected by LIS academics as a research publication outlet, with a SCIMago Journal Ranking position of 82 out of 197 LIS journals ranked (see http://www.scimagojr.com/journalrank.php?category=3309) and placed in the second of four quartiles of journal quality. In comparison, *JOLIS* is ranked at position 22 out of 197 LIS journals and *JELIS* is not ranked in this list. Thus, while *D-Lib* satisfies the criteria for being considered a magazine more often than not, the possible interpretation of this magazine by academics could be one of a higher perceived reliability than a typical magazine, which would explain the smaller similarity value differences classified by STAR′. This perception by academics of *D-Lib* as a quality publication outlet, even though it is classified as a magazine, may indicate that the language used in these articles is closer to the language typically used in LIS journal articles as opposed to more traditional magazine articles. This is not unique to LIS, and similar examples exist in other disciplines, such as the magazine *Communications of the ACM*, which is typically a highly-ranked publication outlet in the field of information systems (Polites & Watson, 2008).

Other factors may also indicate differentiation between journal, magazine, and newsletter articles. The presence of equations and tables is typically thought to be a part of the scientific language (Stiller et al., 2016) and as such might be expected more frequently in journal articles and sought after by graduate students and faculty. Figures or graphics, including animation, are generally considered easier to understand and less technical, and therefore might be found with greater frequency in magazines or newsletters as opposed to journals. Undergraduate

students greatly enjoy graphical interfaces (Landers & Callan, 2011), such as video games, due in part to their immediate feedback and ease of understanding. Therefore, future research is needed to investigate the presence of tables, equations, and figures in articles and their effect on reliance and trust in the information so engendered. While all of these may be seen in all types of publication outlets, their respective presence and ratios may heuristically help improve the STAR′ predictions for magazine articles.

6.1 Utilization

The very high classification accuracy of STAR′ in this exploratory research and consequent documents from additional LIS journals, magazines, and newsletters demonstrates that the STAR′ text mining and text analytics method for classifying publication source type is achievable. These results may be used in a number of ways by students and faculty, as well as by publishers.

Students may be required to find information from particular source types for course assignments or research projects (Cockrell & Jayne, 2002). If the STAR′ phase three classification method is encapsulated in an easily available app, students would be able to rapidly identify the probable publication outlet type to satisfy the requirements of media-type specific assignments. In its current binary classification form, this would require two iterations to fully determine the probable media-type source of information. With the increase in "fake news" from television, to social media, to other web sources, it may be difficult for students to distinguish the reliability of news (Balmas, 2014; Lewis, 2008). When performing research, assuming that academic journals are considered a higher quality and more reliable source of information (Franco, Malhotra, & Simonovits, 2014), both undergraduate and graduate students would be able to use the STAR′ outlet source classifier to estimate the reliability of information gathered from various sources as being comparable to a journal quality article or other levels of reliability associated with magazines and newsletters.

The original STAR algorithm used as the foundation for the development of the STAR′ publication source classification method was reported to be available as an active server page web application. Though the server on which STAR was running is no longer available, a similar web tool could be developed from the STAR′ tool to make it readily available for students to use in evaluating publication outlet source types to satisfy assignment requirements. The tool would also help students in gaining LIS-literacy by being able to rapidly identify if information is coming from a more reliable academic or scientific resource. An alternative application of the STAR′ classification method would be to convert the current C# code into Java to become a Java web application. Such a conversion would enable the method to become platform independent and could also enable

extension of the method to other research domains outside of LIS through a wiki-like interface for creation of domain specific corpuses for each of the three types of article.

Another benefit to students would be identification of words typically used in conveying information to different audiences. The corpuses may be used by the students directly to learn domain-specific vocabulary, which in turn they would be able to use for explaining their own learning and research to prospective employers, who should be more technically literate (i.e., expect and understand the journal-level words), versus friends and family or others who would be expected to be less technically proficient in the domain (i.e., would better understand and receive magazine-level words, intended for a more general audience). Instructors could use the corpuses in a similar manner to instruct information literacy and transference of knowledge to various audiences with differing levels of domain-specific scientific literacy.

Faculty may also utilize the STAR′ method in several ways. The first would be in determining the probable publication outlet source type and consequent reliability of the information reported by students in research or other assignments, especially if this is a condition of satisfactory completion of an assignment. While this evaluation would not necessarily imply the effort put forth by a student, it would be an indicator of a student's information literacy and the possible need for further education in how to critically think about information and acquire appropriate knowledge (Barclay, 2017).

The STAR′ method may also be valuable in helping instructors to teach information literacy, beyond simply identifying students who may need help in this area. An example would be the presentation of several groups of paragraphs of information relevant to the course obtained from various information outlets of varying reliability. After students estimate the probable source, STAR′ would then be used to quantitatively classify the paragraphs into their appropriate sources. An example lesson using a new application integrated with STAR′ is shown in Appendix B.

Students could then use the tool for in-class or homework assignments to reinforce the information literacy concept of information source reliability. Additionally, since it has been shown that students rely primarily on web searches and other web pages as primary information resources (Chu et al., 2016; Griffiths & Brophy, 2005), the STAR′ method could be used in class to compare generic information web pages against sources with known reliability, which should show that some web information pages mimic the high reliability associated with journals, while others have much lower reliability and as such should not be used as information resources to support academic research projects.

Faculty, in performing their own research, may also rely too heavily on search engines for determining which information to utilize in their research (Bonthron et al., 2003). The STAR′ method would enable faculty to rapidly assess the perceived reliability of information sources, as compared to journal articles or other outlet types, and the consequent perception of readers or reviewers in the reliability of the information supporting the faculty member's research. One additional usage of the STAR′ classification method is for faculty to classify their own research prior to publication to assist in determining the appropriate outlet type. Placing research in outlets where the expectations of perceived reliability of information are similar to the comparable reliability of the information presented will promote absorption and utilization of the reported research (Kellogg & Walczak, 2007; Susman & Evered, 1978). Alternatively, if the STAR′ classification is at a level different than the intended publication outlet type, this would indicate that either further work is needed to increase the perceived reliability of the research to an appropriate level for the intended publication outlet type or that the article requires rewording to clarify important points for the intended audience.

Publishers may also be able to utilize the STAR′ outlet type classification method similarly to the proposed use by faculty for determining an appropriate publication outlet type. The increase in online journals has led to pressure for publishers to decrease review times (Jagadish, 2008; Kumar, 2104), which in turn is beneficial to faculty for having their research published more quickly. A preliminary review of submitted articles to a publisher may be quickly performed using the STAR′ outlet type classification method to enable an editor to determine if the perceived reliability of a submitted article is equivalent to the anticipated reliability of articles published in the corresponding outlet. If the submitted article shows sufficient perceived reliability, it can then proceed to other stages of review.

6.2 Limitations

The exploratory research reported in this article serves as a proof of concept for utilizing combined text mining and text analytics for the heuristic classification of a document's publication outlet type. While the results are very promising, having achieved a 100% classification accuracy, there are several limitations to the current research that must be kept in mind. The first is the very small number of documents used for corpus creation. This was necessary due to the decision to treat newsletters as a whole instead of by article (due to the very small article sizes) and the consequent limited availability of newsletter documents over the specified research time period. While evidence for generalizability was obtained by examining articles from two different journals

and a different newsletter, it is unclear how well the current corpuses will permit similarly high classification accuracies when STAR′ is applied to numerous other journals, magazines, and newsletters.

As mentioned previously, future research is needed to continue to develop the LIS journal, magazine, and newsletter corpuses. During this future research, it is likely that some of the regular expression patterns currently contained within a corpus may need to be deleted because they appear within other publication outlet type documents, even with an increased γ value. Additionally, new regular expression patterns will likely be added, creating larger corpuses, which may then require adjusting the *b* value in equation 3 so as not to penalize smaller coverage over larger corpuses. While the corpuses need to be as large as required by adding all regular expression word patterns that satisfy the corpus inclusion heuristics, it is possible that they may still remain relatively small, due to the need for high *tf* and *df* and also the requirement that a term cannot be used more than γ minimum number of times in documents of a different outlet type. Prior research has suggested that even a small-sized corpus may be adequate for classification (Iriondo et al., 2009; Weiss et al., 1999).

All regular expression patterns contained in a corpus are considered equivalent in value for determining classification of a document. However, if the corpus generation technique, based on text mined documents from a known outlet type, places numerous synonymous terms in the corpus, this may negatively affect the classification, depending on author word selection. A possible set of synonymous terms in the journal corpus are the regular expressions learn*, instruct*, educat*, and possibly experienc*. This last word points out some difficulty in determining synonymous terms in that words may have multiple meanings dependent on the domain (Senellart & Blondel, 2004). The effect of coalescing synonymous terms requires further study to determine if their presence is assistive or detractive in the document outlet type classification task.

Another limitation is the necessity of applying STAR′ to a specific domain, in this case LIS. The STAR′ technique is general and a predecessor has shown applicability in the domain of business analytics (Walczak & Kellogg, 2015). The need for applying this methodology to a specific domain is the fact that, as expressed earlier, distinct languages exist for almost every domain of scholastic study and certainly for research within those domains (Luchins, 2007; Teranes, 2013). This means that while the work has already begun for the domain of LIS, the corpus acquisition process must be repeated for each domain in which outlet type classification is useful, including the sciences (though this may require further subdivision into chemistry, biology, physics, geography, etc.), medicine, history or other social sciences, political science, computer science, and so forth.

The last limitation is the choice of using newsletters versus newspapers. The implementation with newspapers faces the same language problem mentioned above, in that the coverage of typical daily newspapers is quite broad and the language used between disciplines may vary significantly, which in turn frustrates the development of a corpus. However, it may be possible to overcome this obstacle if the newspaper articles used for both corpus development and later classification are all from a common domain, such as cybersecurity, or population health, since a common language may exist for reporting phenomena within these domains at the level of a newspaper article targeted for consumption by the general public.

6.3 Future Research

The corpuses for the LIS domain need to be further extended. The *D-Lib* magazine used for development of the magazine corpus indicated some difficulties in distinguishing between articles published in *D-Lib* versus journals, possibly due to the high academic reliance on this magazine as a quality publication outlet. While the journal and newsletter corpuses should be extended by incorporating regular expression word patterns from new journals and newsletters, it is critical that magazines which are viewed academically as being more similar to traditional magazine outlets be used to develop the magazine corpus. This might mean that the *DLib* articles would end up being more frequently classified as journal articles, but this could be a more accurate representation of the level of and reliability of the knowledge conveyed in this particular magazine.

Future research is needed to extend the STAR′ text mining and text analytics-based heuristic similarity-based classification to other domains beyond LIS. This will require development of journal, magazine, and newsletter corpuses within each new domain being researched. Prior research has already shown a requirement for classifying publication outlet types in the science domains (Klosterman, Sadler, & Brown, 2012; Tseng et al., 2010) and in general for undergraduate education (Cockrell & Jayne, 2002). As this research progresses, it would be worthwhile to investigate the development of a cloud-based resource or web resource to house collections of corpuses verified in various different domains so that educators and researchers would have access to these verified corpuses to perform document classifications, without having to recreate their own corpuses.

Phases one and three of the STAR′ method are automated as C# programs. However, the phase two word pattern analysis across documents and classes is still largely manual. Some automation has been performed as spreadsheet macros, but the entire process requires future research to investigate if automation of

phase two and subsequent corpus creation is possible and also to perform such automation if deemed possible. Determining if the same word(s) exist in multiple documents is fairly straightforward (e.g., any search tool in word processors), but clustering like patterns together for creation of the regular expressions to be used in a corpus is more problematic. Other research could examine the benefit and usability to students achieved from encapsulating the STAR′ classifier into a phone app for improved ease of use and perceived usefulness (Calisir & Calisir, 2004).

The STAR′ document classifier is in essence a binary classifier since it can only distinguish between two different classes of document outlet types at a time. Although different from strict binary classifiers because it provides a real number value assessing a document's similarity to the other documents in that class used to produce the class corpus, the end result is a correct or incorrect classification into one of two groups or a decision that a classification cannot be made due to lack of sufficient differences between a document and the corpuses for the two classes. Future research should examine the improvement of the STAR′ method to become an *n*-ary classifier, being able to classify a document into one of *n* distinct classes, which for the research reported in this article *n* would take on a value of three for journal article, magazine article, or newsletter. Corpuses for each of the *n* classes must be developed and provided to the phase three STAR′ classifier. The similarity ranking of a document to a class uses a scale of real number values with the class designated by the sign (negative or positive) of the heuristic classification value. The similarity encoding will need to be augmented to be able to now account for the *n* distinct classes, as opposed to just two classes of document outlet type. Additional classes of outlet type could include newspaper articles, web pages, corporate documents, and marketing materials, among others. Since *n* possible internal values will need to be compared, it may occur that a document is highly similar to multiple types of outlets (Frigui & Nasraoui, 2004), and a method for determining a correct classification given these *n* values will need to be developed, and may possibly utilize fuzzy algorithms (Chiclana, Herrera, & Herrera-Viedma, 2001; Herrera & Herrera-Viedma, 2000).

7.0 Conclusions

> "Media are evolving at extraordinary rates, and individuals in modern society are accessing, using, influencing and being influenced by media in ways that have not been previously observed. We believe that even casual observers of schools, education systems, and educational research would readily agree that these areas have not held pace with the rapidly changing landscape of media and how the new media landscapes impact

> teaching and learning (Klosterman, Salder, & Brown, 2012, p. 52)."

Information literacy is a critical skill for student success at the university level (Laubersheimer, Ryan, & Champaign, 2016). Classification of textual media will improve information literacy through evaluation of the reliability of reported research information. Text analytics has previously been shown to be useful for performing identification of concepts, text summarization, and text classification of various media documents. Examples include identifying scientific concepts from new articles (Tseng et al., 2010), grouping documents by specific research findings within a specific domain (Bichindaritz & Akkineni, 2006), and determining theoretical versus applied research methodology of articles (Walczak & Kellogg, 2015).

The research reported in this article seeks to develop and prove the efficacy of using a text mining and text analytics tool to automatically classify the publication outlet type of a research article with respect to a journal article, magazine article, or newsletter. The method also shows a heuristic method for mining text words from documents and using the mined text dictionaries from a collection of documents to heuristically determine an appropriate corpus.

The STAR′ publication outlet type classification methodology, including corpus development through text mining and heuristic analysis of word patterns for corpus development, has been shown to accurately classify all LIS journal and magazine articles and full newsletters across three different LIS journals, one LIS magazine, and two LIS newsletters. Three research hypotheses were proposed with each concerning the ability of STAR′ to accurately identify articles originating from a specific type of publication outlet in the LIS domain. The performance of STAR′ supported each of the three hypotheses, though the magazine side of hypotheses H_2 and H_3 is only partially supported (70% for H2 and 87% for H3). While this research is exploratory, it does demonstrate the efficacy of using text analytics to classify an article's publication outlet type and the need for further research to validate the proposed method across a larger contingent of LIS journals, magazines, and newsletters, as well as extending the research to other research fields (e.g., medicine, sciences, and social sciences) where disambiguation of outlet source may be beneficial.

The proposed STAR′ methodology may be used by students to improve perceptions of the reliability of information used in completing their studies. Faculty may use the method for analyzing reported student information sources, identifying information literacy shortcomings in their students, analyzing the potential reliability of published research for use in their own research, and

evaluating the potential fit of their pre-publication status research and potential publication outlets.

The STAR′ methodology proposed may be generalized to classify other properties or aspects of research articles. Specifically, how to evaluate words or regular expression patterns for inclusion in a domain corpus and the heuristics used in applying a corpus to a research article for classification are not specific to classifying a research article's publication outlet type or associated reliability. These methods may be generalized to other classification problems involving the analysis and classification of textual documents.

Author Biography

Steven Walczak is an Associate Professor in the School of Information and for the Florida Center for Cybersecurity at the University of South Florida in Tampa, Florida. Dr. Walczak earned his Ph.D. from the University of Florida, M.S. in computer science and B.S. in mathematics from the Johns Hopkins University and the Pennsylvania State University respectively. His primary research foci are applications of artificial intelligence and health information systems. Dr. Walczak is Editor-in-Chief for the *Journal of Organizational and End User Computing* and serves as an AE for the *International Journal of Healthcare Information Systems and Informatics*.

References

Aggarwal, C. C., & Zhai, C. X. (2012). An introduction to text mining. In C. C. Aggarwal & C. X. Zhai (Eds.), *Mining text data* (pp. 1-10). New York: Springer. Retrieved from https://pdfs.semanticscholar.org/6806/588dfc2633c3cd86ab3faba2a44a8e4ceb63.pdf#page=13

Balmas, M. (2014). When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research, 41*(3), 430-454. Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/0093650212453600

Baoli, L., Qin, L., & Shiwen, Y. (2004). An adaptive k-nearest neighbor text categorization strategy. *ACM Transactions on Asian Language Information Processing*, *3*(4), 215-226. Retrieved from http://dl.acm.org/citation.cfm?id=1039623

Barclay, D. A. (2017, January 4). The challenge facing libraries in an era of fake news. *The Conversation*. Retrieved from https://theconversation.com/the-challenge-facing-libraries-in-an-era-of-fake-news-70828

Bichindaritz, I., & Akkineni, S. (2006). Concept mining for indexing medical literature. *Engineering Applications of Artificial Intelligence, 19*(4), 411-417. Retrieved from http://dx.doi.org/10.1016/j.engappai.2006.01.009

Bonthron, K., Urquhart, C., Thomas, R., Armstrong, C., Ellis, D., Everitt, J., Fenton, R., Lonsdale, R., McDermott, E., Morris, H., & Phillips, R. (2003). Trends in use of electronic journals in higher education in the UK: Views of academic staff and students. *D-Lib Magazine, 9*(6), 16. Retrieved from http://www.dlib.org/dlib/june03/urquhart/06urquhart.html

Bragge, J., Thavikulwat, P., & Töyli, J. (2010). Profiling 40 years of research in Simulation & Gaming. *Simulation & Gaming, 41*(6), 869-897. Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/1046878110387539

Bui, D. D. A., Del Fiol, G., & Jonnalagadda, S. (2016). PDF text classification to leverage information extraction from publication reports. *Journal of Biomedical Informatics, 61*, 141-148. Retrieved from http://www.sciencedirect.com/science/journal/15320464/61

Calisir, F., & Calisir, F. (2004). The relation of interface usability characteristics, perceived usefulness, and perceived ease of use to end-user satisfaction with enterprise resource planning (ERP) systems. *Computers in Human Behavior, 20*(4), 505-515. Retrieved from http://www.sciencedirect.com/science/journal/07475632/20/4

Chiclana, F., Herrera, F., & Herrera-Viedma, E. (2001). Integrating multiplicative preference relations in a multipurpose decision-making model based on fuzzy preference relations. *Fuzzy Sets and Systems, 122*(2), 277-291. Retrieved from http://www.sciencedirect.com/science/journal/01650114/122/2

Chu, S. K., Lau, W. W., Chu, D. S., Lee, C. W., & Chan, L. L. (2016). Media awareness among Hong Kong primary students. *Journal of Librarianship and Information Science, 48*(1), 90-104. Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/0961000614551448

Cockrell, B. J., & Jayne, E. A. (2002). How do I find an article? Insights from a web usability study. *The Journal of Academic Librarianship, 28*(3), 122-132. Retrieved from http://www.sciencedirect.com/science/article/pii/S0099133302002793

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, *6*(1), 57-71.

Conway, M. (2010). Mining a corpus of biographical texts using keywords. *Literary and Linguistic Computing*, *25*(1), 23-35. Retrieved from https://academic.oup.com/dsh/article-abstract/25/1/23/926040/Mining-a-corpus-of-biographical-texts-using

Crain, S. P., Zhou, K., Yang, S. H., & Zha, H. (2012). Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In C. C. Aggarwal & C. X. Zhai (Eds.), *Mining text data* (pp. 129-161). New York: Springer.

Davis, P. M., & Cohen, S. A. (2001). The effect of the Web on undergraduate citation behavior 1996–1999. *Journal of the American Society for Information Science and Technology*, *52*(4), 309-314. Retrieved from https://ecommons.cornell.edu/bitstream/handle/1813/2557/52.4davis.pdf?sequence=1

Dietrich, D., Heller, B., & Yang, B. (2015). *Data science and big data analytics*. Indianapolis: Wiley.

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76-82. Retrieved from http://cacm.acm.org/magazines/2006/9/5835-tapping-the-power-of-text-mining/abstract

Fitzgerald, M. (2012). *Introducing regular expressions*. Sebastopol: O'Reilly Media.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502-1505. Retrieved from http://www.law.nyu.edu/sites/default/files/upload_documents/September 9 Neil Malhotra.pdf

Frigui, H., & Nasraoui, O. (2004). Simultaneous clustering and dynamic keyword weighting for text documents. In M. W. Berry (Ed.), *Survey of text mining* (pp. 45-72). New York: Springer. Retrieved from https://www.researchgate.net/profile/Olfa_Nasraoui/publication/240242284_Simultaneaous_Clustering_and_Dynamic_Keyword_Weighting_for_Text_Douments/links/0c960531a4dcab76f0000000.pdf

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. Retrieved from http://www.sciencedirect.com/science/article/pii/S0268401214001066

Goutte, C. (2008). A probabilistic model for fast and confident categorisation of textual documents. In M. W. Berry & M. Castellanos (Eds.), *Survey of text mining II* (pp. 189-202). London: Springer. Retrieved from https://www.researchgate.net/profile/Cyril_Goutte/publication/44070027_A_Probabilistic_Model_for_Fast_and_Confident_Categorisation_of_Textual_Documents/links/0fcfd50b12facb6e5e000000.pdf

Griffiths, J. R., & Brophy, P. (2005). Student searching behavior and the Web: Use of academic resources and Google. *Library Trends*, *53*(4), 539-554. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/1749/Griffiths539554.pdf?sequence=2

Head, A. J., & Wihbey, J. (2014, July 17). At sea in a deluge of data. *The Chronicle of Higher Education*, 3. Retrieved from http://chronicle.com/article/At-Sea-in-a-Deluge-of-Data/147477/

Herrera, F., & Herrera-Viedma, E. (2000). Linguistic decision analysis: Steps for solving decision problems under linguistic information. *Fuzzy Sets and systems*, *115*(1), 67-82. Retrieved from http://www.sciencedirect.com/science/journal/01650114/115/1

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum*, *20*(1), 19-62. Retrieved from http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635-650. Retrieved from http://www.radford.edu/~jaspelme/443/spring-2007/Articles/Hovland_n_Weiss_1951_sleeper-effect.pdf

Iriondo, I., Planet, S., Socoró, J. C., Martínez, E., Alías, F., & Monzo, C. (2009). Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification. *Speech Communication*, *51*(9), 744-758. Retrieved from https://hal.archives-ouvertes.fr/hal-00550285/document

Isa, D., Kallimani, V. P., & Lee, L. H. (2009). Using the self organizing map for clustering of text documents. *Expert Systems with Applications*, *36*(5), 9584-

9591. Retrieved from https://www.researchgate.net/profile/Lam_Hong_Lee/publication/223420632_Using_the_self_organizing_map_for_clustering_of_text_documents/links/00b7d536ad0f25b3c8000000.pdf

Jagadish, H. V. (2008). The conference reviewing crisis and a proposed solution. *ACM SIGMOD Record*, *37*(3), 40-45. Retrieved from https://sigmodrecord.org/publications/sigmodRecord/0809/p40.jagadish.pdf

Jamieson, S. (2016). What the citation project tells us about information literacy in college composition. In B. J. D'Angelo, S. Jamieson, B. Maid, & J. R. Walker (Eds.), *Information literacy: Research and collaboration across disciplines* (pp. 115-138). Fort Collins, CO: WAC Clearinghouse and University Press of Colorado. Retrieved from https://wac.colostate.edu/books/infolit/collection.epub

Kellogg, D. L., & Walczak, S. (2007). Nurse scheduling: From academia to implementation or not? *Interfaces*, *37*(4), 355-369. http://dx.doi.org/10.1287/inte.1070.0291

Kim, S. B., Han, K. S., Rim, H. C., & Myaeng, S. H. (2006). Some effective techniques for Naive Bayes text classification. *IEEE Transactions on Knowledge and Data Engineering*, *18*(11), 1457-1466. Retrieved from http://koasas.kaist.ac.kr/bitstream/10203/16860/1/some effective techniques for naive bayes text classification.pdf

Kimble, J. (2013). You think the law requires legalese. *Michigan Bar Journal*, *92*(11), 48-50. Retrieved from https://www.michbar.org/file/barjournal/article/documents/pdf4article2296.pdf

Kissel, F., Wininger, M. R., Weeden, S. R., Wittberg, P. A., Halverson, R. S., Lacy, M., & Huisman, R. K. (2016). Bridging the gaps: Collaborating in a faculty and librarian community of practice on information literacy. In B. J. D'Angelo, S. Jamieson, B. Maid, & J. R. Walker (Eds.), *Information literacy: Research and collaboration across disciplines* (pp. 411-428). Fort Collins, CO: WAC Clearinghouse and University Press of Colorado. Retrieved from https://wac.colostate.edu/books/infolit/collection.epub

Klosterman, M. L., Sadler, T. D., & Brown, J. (2012). Science teachers' use of mass media to address socio-scientific and sustainability issues. *Research in Science Education*, *42*(1), 51-74. Retrieved from http://link.springer.com/article/10.1007/s11165-011-9256-z

Korhonen, A., Séaghdha, D. O., Silins, I., Sun, L., Högberg, J., & Stenius, U. (2012). Text mining for literature review and knowledge discovery in cancer risk assessment and research. *PloS one*, *7*(4), e33427. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033427

Kumar, M. N. (2014). Review of the ethics and etiquettes of time management of manuscript peer review. *Journal of Academic Ethics*, *12*(4), 333-346. Retrieved from http://link.springer.com/article/ 10.1007/s10805-014-9220-4

Landers, R. N., & Callan, R. C. (2011). Casual social games as serious games: The psychology of gamification in undergraduate education and employee training. In M. Ma, A. Oikonomou, &! L. C. Jain (Eds.), *Serious games and edutainment applications* (pp. 399-423). London: Springer.

Laskin, M., & Haller, C. R. (2016). Up the mountain without a trail: Helping students use source networks to find their way. In B. J. D'Angelo, S. Jamieson, B. Maid, & J. R. Walker (Eds.), *Information literacy: Research and collaboration across disciplines* (pp. 237-256). Fort Collins, CO: WAC Clearinghouse and University Press of Colorado. Retrieved from https://wac.colostate.edu/books/infolit/collection.epub

Laubersheimer, J., Ryan, D., & Champaign, J. (2016). InfoSkills2Go: Using badges and gamification to teach information literacy skills and concepts to college-bound high school students. *Journal of Library Administration*, *56*(8), 924-938. Retrieved from http://www.tandfonline.com/doi/full/10.1080/01930826.2015.1123588

Lewis, S. C. (2008). Where young adults intend to get news in five years. *Newspaper Research Journal*, *29*(4), 36-52.

Lloyd, A. (2005). Information literacy: different contexts, different concepts, different truths? *Journal of Librarianship and Information Science*, *37*(2), 82-88. Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/0961000605055355

Luchins, D. J. (2007). Corporate speak and the psychiatric profession. *Administration and Policy in Mental Health and Mental Health Services Research*, *34*(4), 421-423. Retrieved from http://link.springer.com/article/10.1007%2Fs10488-007-0112-4?LI=true

Ma, J., Xu, W., Sun, Y. H., Turban, E., Wang, S., & Liu, O. (2012). An ontology-based text-mining method to cluster proposals for research project selection. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and*

Humans, *42*(3), 784-790. Retrieved from https://pdfs.semanticscholar.org/37c3/259b5e5652c4c3019fccce15708942c9732c.pdf

McCune, J. C. (1999). Do you speak computerese? *Management Review*, *88*(2), 10-12. Retrieved from http://search.proquest.com/openview/f0af24e4bbf34a4ec1b868cbd2fe76fb/1?pq-origsite=gscholar&cbl=41493

Metzger, M. J., Flanagin, A. J., & Zwarun, L. (2003). College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education*, *41*(3), 271-290. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.7139&rep=rep1&type=pdf

Mogge, D. (1999). Seven years of tracking electronic publishing: The ARL Directory of Electronic Journals, Newsletters and Academic Discussion Lists. *Library Hi Tech*, *17*(1), 17-25. Retrieved from http://www.emeraldinsight.com/doi/pdfplus/10.1108/07378839910267154

Polites, G. L., & Watson, R. T. (2008). The centrality and prestige of CACM. *Communications of the ACM*, *51*(1), 95-100.Retrieved from http://cacm.acm.org/magazines/2008/1/5490-the-centrality-and-prestige-of-cacm/abstract

Reed, K. L. (1999). Mapping the literature of occupational therapy. *Bulletin of the Medical Library Association*, *87*(3), 298-304. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC226589/pdf/mlab00088-0068.pdf

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503-520. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.7340&type=pdf&rep=rep1

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1-47. Retrieved from https://arxiv.org/pdf/cs/0110053.pdf

Senellart, P. P., & Blondel, V. D. (2004). Automatic discovery of similar words. In M. W. Berry (Ed.), *Survey of text mining* (pp. 25-43). New York: Springer.

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, *41*(5), 721-732. Retrieved from https://www.researchgate.net/profile/Annette_Upmeier_zu_Belzen/publication/300015480_Assessing_scientific_reasoning_a_comprehensive_evaluation_of_item_features_that_affect_item_difficulty/links/5708216d08aea66081332247.pdf

Sun, Y., Deng, H., & Han, J. (2012). Probabilistic models for text mining. In C. C. Aggarwal & C. X. Zhai (Eds.), *Mining text data* (pp. 259-295). New York: Springer.

Susman, G. I, & Evered, R. D. (1978). An assessment of the scientific merits of action research. *Administrative Science Quarterly*, *23*(4), 582-603. Retrieved from http://dx.doi.org/10.2307/2392581

Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research*, *1*(1), 77-100. Retrieved from http://mmr.sagepub.com/cgi/content/abstract/1/1/77

Teranes, P. S. (2013). Make it as simple as you can. *Michigan Bar Journal*, *92*(11), 52-53. Retrieved from https://www.michbar.org/file/barjournal/article/documents/pdf4article2297.pdf

Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification. *International Journal of Software Engineering and Its Applications*, *5*(3), 37-46. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.643.6611&rep=rep1&type=pdf

Tseng, Y. H., Chang, C. Y., Rundgren, S. N. C., & Rundgren, C. J. (2010). Mining concept maps from news stories for measuring civic scientific literacy in media. *Computers & Education*, *55*(1), 165-177. Retrieved from http://dx.doi.org/10.1016/j.compedu.2010.01.002

Walczak, S., & Kellogg, D.L. (2015). A heuristic text analytic approach for classifying research articles. *Intelligent Information Management*, *7*(1), 7-21. Retrieved from http://file.scirp.org/Html/2-8701325_53509.htm

Weiss, S. M., Apte, C., Damerau, F. J., Johnson, D. E., Oles, F. J., Goetz, T., & Hampp, T. (1999). Maximizing text-mining performance. *IEEE Intelligent Systems*, *14*(4), 63-69.

Wiebe, T. (2016). The information literacy imperative in higher education. *Liberal Education*, *102*(1), 52–55. Retrieved from http://digitalcommons.hope.edu/cgi/viewcontent.cgi?article=2521&context=faculty_publications

Worsley, A. (1989). Perceived reliability of sources of health information. *Health Education Research*, *4*(3), 367-376. Retrieved from http://her.oxfordjournals.org/content/4/3/367.short

Yamamoto, M., & Church, K. W. (2001). Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, *27*(1), 1-30. Retrieved from http://www.mitpressjournals.org/doi/pdfplus/10.1162/089120101300346787

Yancey, K. B. (2016). Creating and exploring new worlds: Web 2.0 information literacy and the ways we know. In B. J. D'Angelo, S. Jamieson, B. Maid, & J. R. Walker (Eds.), *Information literacy: Research and collaboration across disciplines* (pp. 77-91). Fort Collins, CO: WAC Clearinghouse and University Press of Colorado. Retrieved from https://wac.colostate.edu/books/infolit/collection.epub

Yao, Q. (2009). An evidence of frame building: Analyzing the correlations among the frames in Sierra Club newsletters, national newspapers, and regional newspapers. *Public Relations Review*, *35*(2), 130-132.

Young, M. E., Norman, G. R., and Humphreys, K. R. (2008). The role of medical language in changing public perceptions of illness. *PLoS One*, *3*(12), e3875. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003875

Appendix A

The corpuses created for LIS journal articles, LIS magazine articles, and LIS newsletters are shown below.

Journal Corpus	**Magazine Corpus**	**Newsletter Corpus**
(26 regular expressions)	(21 regular expressions)	(23 regular expressions)
respond*	abilit*	health
profession*	account	medic*
skill*	analy*	fhsla
learn*	branch*	member*
experienc*	content	science?
research*	countr*	meeting
course?	criteri*	annual*
survey?	data	resource?
instruct*	digital	committee?
provid*	evidenc*	study
educat*	identif*	facult*
develop*	indicat*	servic*
curricul*	institution?	award?
communicat*	manag*	year?
platform?	metadata	area?
service*	outcom*	collection?
question*	repositor*	submit*
source?	review*	attend*
online	social	educat*
study	standard?	president
task*	system?	team*
field?		public
program*		circulat*
understand*		
response?		
tool?		

The * symbol is a wildcard and represents matching from zero to an infinite

number of other alphabetic characters. The ? symbol is another type of wildcard and represents matching either zero or one additional alphabetic character. A space at the front of an expression indicates that this expression must appear at the beginning of the word pattern, which occurs for the newsletter pattern " fhsla" and the magazine pattern " social ". A space at the end of an expression (which cannot be seen) indicates that nothing may follow the regular expression, which occurs only for the magazine pattern " social ". All whitespace characters (space, tab, newline) or numeric characters end the current string under consideration as a regular expression match.

Appendix B

A lesson on information literacy using a C# program that invokes STAR′ is demonstrated below.

The instructor needs to pre-select examples of information obtained from journals, magazines, newspapers or newsletters, and general informational websites. These examples are loaded into a small application, written in C#, on information source literacy prior to the class.

In the class, the instructor should first illustrate the differences between journal articles, magazine articles, newspaper articles, and general informational websites. During this demonstration, the quality and reliability of each publication source should also be illustrated and emphasized, perhaps using a group discussion to evoke student perceptions on each publication source and then explanations for why the various sources have differing perceptions of reliability.

After students have discussed the differences between the various article types the program can be executed (see Figure B1). The instructor clicks on the next example and an example for each of two different source types is selected from the pre-loaded examples and displayed to the students as shown in Figure B2. When the program displays the information content from two contrasting publication sources, this may be used for the preliminary discussion illustrating differences between media outlets. After the B2 screen is shown for the application portion of the class, the instructor either selects a student to answer the question posed for the example on the left-side of the screen or enables a class discussion to select the expected source and clicks the corresponding button, as shown in Figure B3. The process is repeated for the example on the right-side of the screen, as shown in Figure B4. At this time, the instructor can use the two examples to further illustrate the differences in quality and reliability between the two displayed types. Additional examples may be randomly chosen from the pre-

loaded examples, by clicking on the "Next comparison" button, with the program always ensuring that two distinct publication source types are shown together.

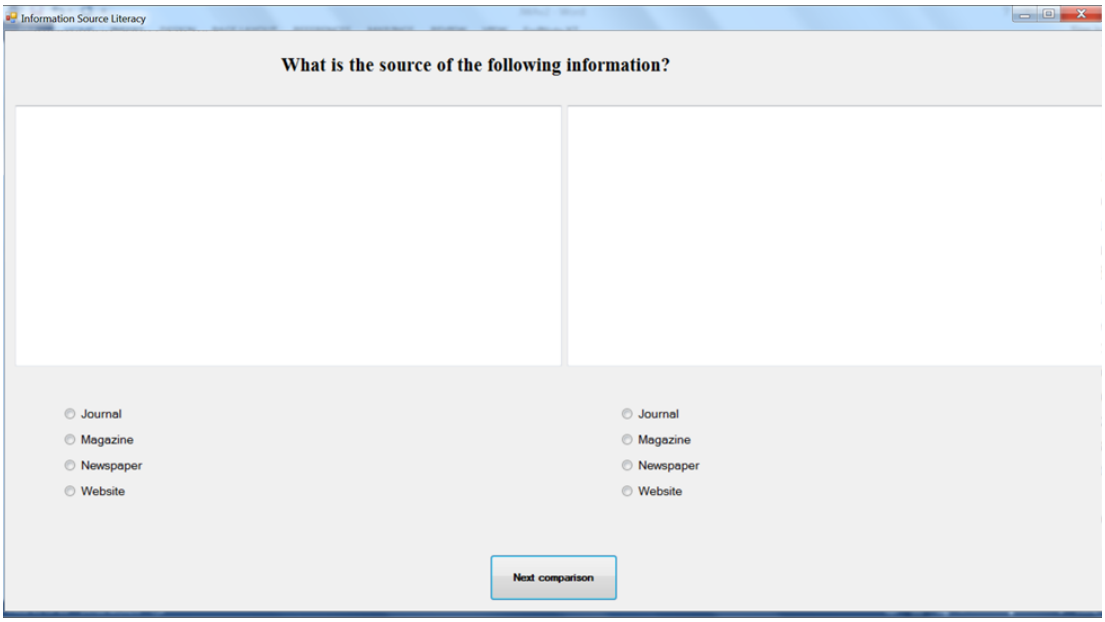


Figure B1. Initial screen of Information Source Literacy classroom teaching application.

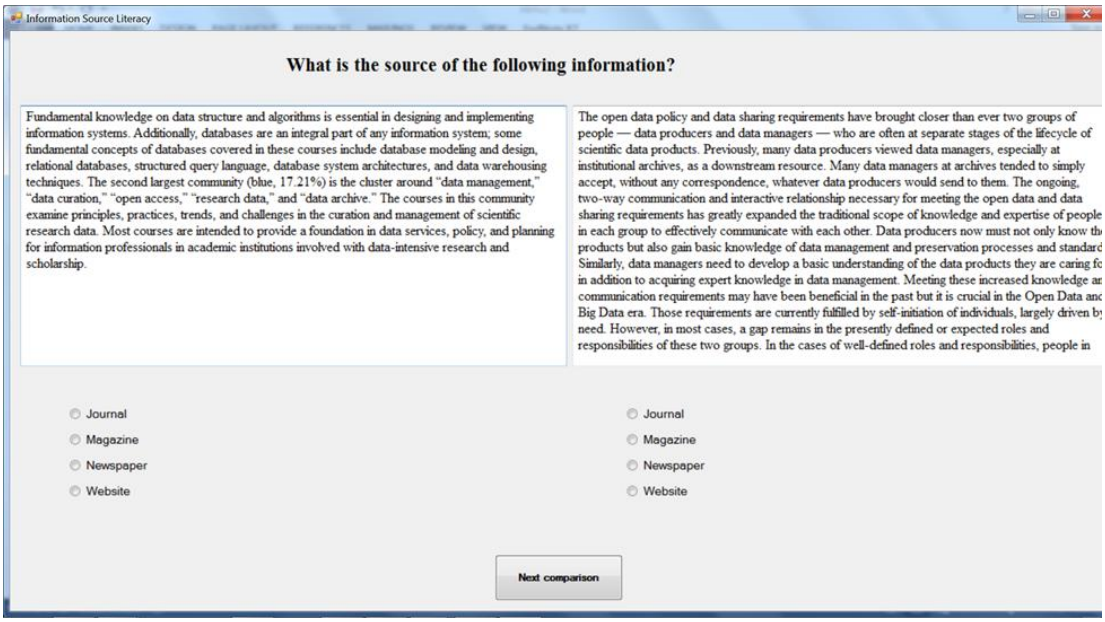


Figure B2. Information Source Literacy classroom teaching application showing comparison of information from two publication sources.

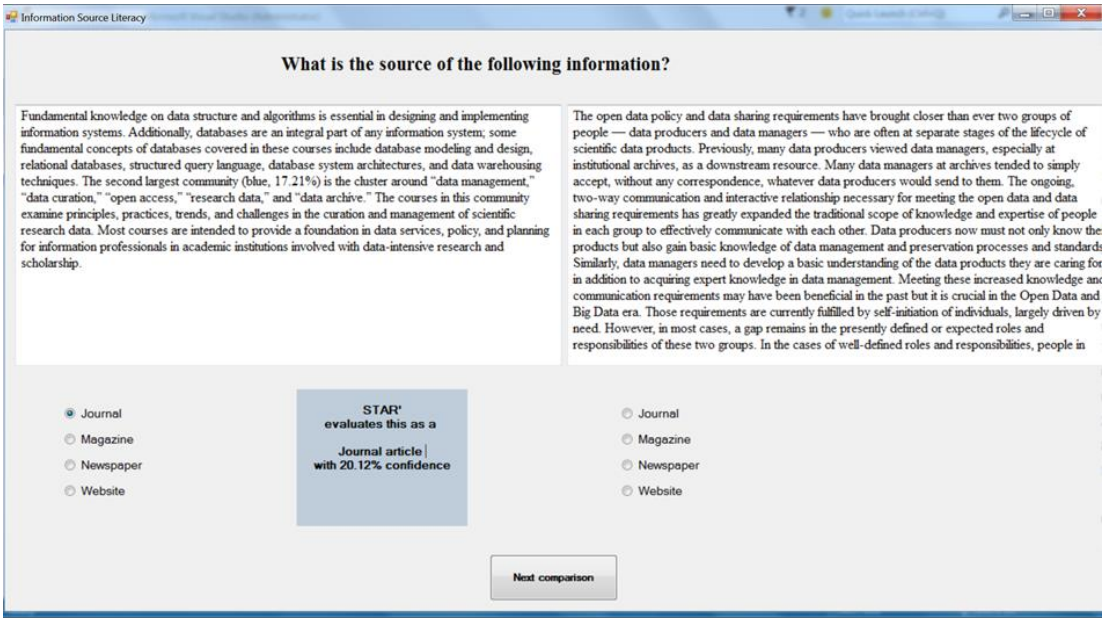


Figure B3. Information Source Literacy classroom teaching application showing selection of first information source type.

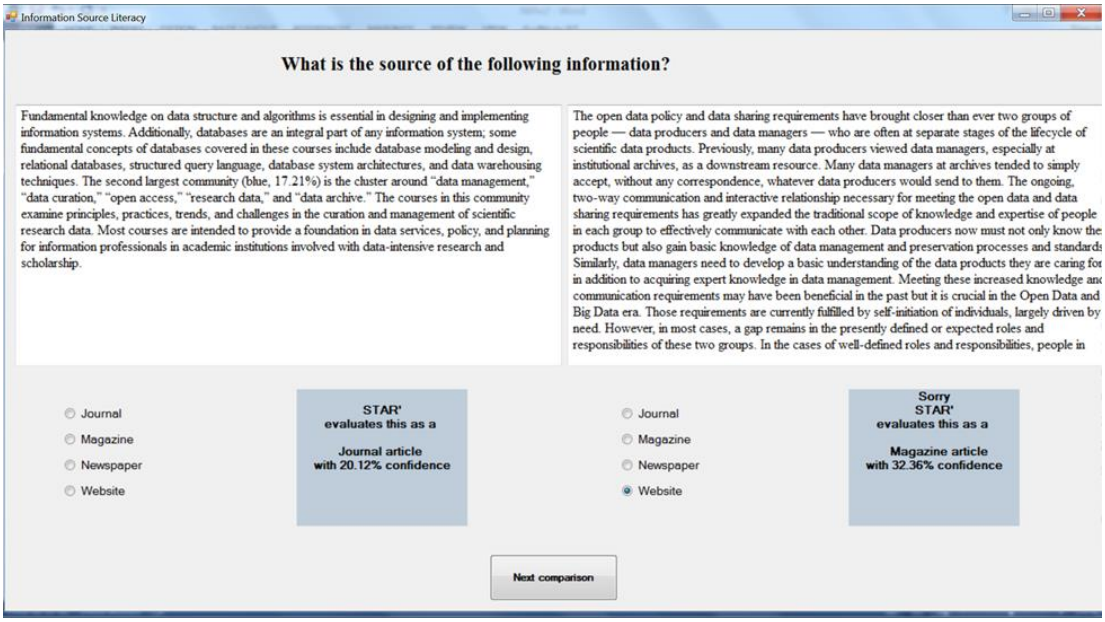


Figure B4. Information Source Literacy classroom teaching application showing incorrect selection of second information source type.