International Journal of Accounting Information Systems 38 (2020) 100469

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

journal homepage: www.elsevier.com/locate/accinf

# An ontological artifact for classifying social media: Text mining analysis for financial data☆

Zamil Alzamil [a], Deniz Appelbaum [b,*], Robert Nehmer [c]

[a] Computer Science Department, College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia
[b] Accounting and Finance Department, Feliciano School of Business, Montclair State University, Montclair, NJ, USA
[c] Accounting and Finance Department, School of Business Administration, Oakland University, Rochester, MI, USA

ARTICLE INFO

Available online 31 July 2020

*Keywords:*
FIBO
Ontology
Social media
Frames and slots
Municipal bonds

In this paper we utilize a structured natural language processing implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from the unstructured textual data of the social media platform Twitter regarding financial and budget information in the public sector, namely the two public-private agencies of the Port Authority of NY and NJ (PANYNJ), and the NY Metropolitan Transportation Agency (MTA). This research initiative uses the Design Science Research (DSR) perspective to develop an artifact to classify tweets as being either relevant to financial bonds or not. We apply a frame and slot approach from the artificial intelligence and natural language processing literature to operationalize this artifact. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts. We show that FIBO grammar can be used to mine semantic meaning from unstructured textual data and that it provides a nuanced representation of structured financial data. With this artifact, social media such as Twitter may be accessed for the knowledge that its text contains about financial concepts using the FIBO ontology. This process is anticipated to be of interest to bond issuers, regulators, analysts, investors, and academics. It may also be extended towards other financial domains such as securities, derivatives, commodities, and banking that relate to FIBO ontologies, as well as more generally to develop a structured knowledge representation of unstructured data through the application of an ontology.


## 1. Introduction

In this paper we utilize a natural language processing implementation of the Financial Industry Business Ontology (FIBO) to extract financial information from the social media platform Twitter regarding financial and budget information in the public sector, namely the collective public-private agencies of the Port Authority of NY and NJ (PANYNJ), and the NY Metropolitan Transportation Agency (MTA). This research initiative is approached from a Design Science Research (DSR) perspective to develop an artifact to classify tweets as being either about financial bonds or not. We apply a frame and slot methodology from the artificial intelligence and natural language processing literature to operationalize the artifact using the FIBO ontology in the public sector/municipalities business context. FIBO is part of the Enterprise Data Management Council (EDMC) and Object Management

☆ This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We thank the referees and participants at the 2019 UW CISA 11th Biennial Symposium on Information Integrity and Information Systems Assurance for helpful comments, especially the discussants William Ives and Ryan Baxter.

* Corresponding author at: 1 Normal Avenue, SBUS #387, Montclair, NJ 07043, USA.
*E-mail addresses:* z.alzamil@mu.edu.sa, (Z. Alzamil), appelbaumd@montclair.edu, (D. Appelbaum), nehmer@oakland.edu. (R. Nehmer).

https://doi.org/10.1016/j.accinf.2020.100469
1467-0895/

Group (OMG) family of specifications. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts. One contribution of this paper is the recognition that the FIBO structure provides a grammar of financial concepts which can be used to classify social media for knowledge representation. Knowledge representation is the process of capturing many different representations of concepts which is facilitated by the use of an ontology or other methods. Previous research explores non-financial ontologies for knowledge representation and FIBO ontologies for sentiment analysis in social media, but research has yet to discuss financial ontologies for knowledge representation extraction from social media. We show that this grammar can be used to mine semantic meaning[1] from unstructured textual data. Twitter streams are monitored and frames derived from FIBO and key words. The ability of the FIBO frames to detect semantic meaning in tweets is compared with naïve key word analysis and by determining the number of false positives classified from the Twitter stream. Using FIBO frames, constituent semantic structures can be uncovered to predict reactions to policies and programs and do other environmental scanning more quickly than by following the feeds manually. With this artifact, social media such as Twitter may be accessed for the knowledge that its utterances contain about financial concepts at many levels.

### 1.1. Design science

Capturing and examining unstructured Twitter data requires a structured design process for analysis. The process utilized by this research is Design Science Research (DSR). Design Science Research aims to solve problems through the design and creation of artifacts (Geerts et al., 2013). There are four types of artifacts (March and Smith, 1995): Constructs (e.g. vocabulary of a domain), Models (a set of proposals that reflects relationships among constructs), Methods (steps to perform a task), and Instantiations (the application of an artifact). To be considered as research, the artifact should be applied within its relevant field and address an important unsolved problem with an innovative solution (Hevner et al., 2004). Design Science Research in the accounting field is strongly based on such artifacts – consider Activity Based Costing, the REA Enterprise Systems (McCarthy, 1982), the COSO Framework, the XBRL reporting taxonomy, to name a few innovations which were designed and have been implemented to address important accounting problems (Geerts et al., 2013). In agreement with McCarthy (2012),[2] this paper follows the DSR approach introduced by Hevner et al. (2004), Stutzman (2007), and which has been further clarified by Geerts (2011). Following the proposal of Geerts (2011), this paper applies DSR for categorization of the methodology into six summary processes:

1. Problem identification and motivation
2. Definition of the objectives of a solution
3. Design and development of an artifact which meets (some of) the solution objectives
4. Demonstration of the solution
5. Evaluation of the solution
6. Communication of the problem and the solution (usually an academic paper).

### 1.2. Twitter feeds and ontologies

The financial woes of the MTA and PANYNJ are becoming likely subjects for social media feeds, such as Twitter, and these tweets could serve as a potentially rich data source about a financial topic. Stakeholders at many levels may want to avail themselves of the potential financial information residing within these social media sources. That is, does the revenue number reported in the financial statements represent an efficient and effective collection of all subway and bus fares? Or, are there inefficiencies of the system that would imply that all potential revenue (tickets, tokens) is not being collected? What better source exists for such information than a social media platform like Twitter, where users may post at whim regarding their observations and opinions? This paper attempts to provide structure to this task of knowledge representation by mining Twitter data feeds using FIBO ontology terms that pertain to these quasi-public PANYNJ and MTA funds. What follows is a discussion of prior literature on ontologies and Twitter analysis, a description of the research, a description of the implementation, and concluding thoughts.

This study contributes to the accounting field by addressing the challenge and need of firm management, auditors, and regulators to deal systematically with unstructured social media that is relevant to their interests. Academically, we are not aware of any other study that has applied a formal business ontology for knowledge representation from social feeds, such as Twitter. This process allows the tweets in the feed to be classified as either relevant to an environmental scan that relates to a financial concept or not. We envision a situation where either firm management or their auditors are looking to scan the firm's environment for potential risks. Social media provides a rich potential source of environmental signals but also carries a lot of noise. The FIBO ontology was built using domain experts whose financial expertise has been captured in the ontology. Therefore, using such an ontology facilitates the identification and understanding of the nuanced threads in Twitter that pertain to the knowledge extraction

[1] According to Wikipedia: "In linguistics, semantics is the subfield that is devoted to the study of meaning, as inherent at the levels of words, phrases, sentences, and larger units of discourse (termed texts, or narratives). The study of semantics is also closely linked to the subjects of representation, reference and denotation. The basic study of semantics is oriented to the examination of the meaning of signs, and the study of relations between different linguistic units…A key concern is how meaning attaches to larger chunks of text, possibly as a result of the composition from smaller units of meaning. Traditionally, semantics has included the study of sense and denotative reference, truth conditions, argument structure, thematic roles, discourse analysis, and the linkage of all of these to syntax". https://en.wikipedia.org/wiki/Semantics.

[2] McCarthy (2012) states: "Accounting has become more dogmatic and dominated by the conventional wisdom of just a few groups" and "Much of the accounting research prior to 1970 was intrinsically normative, arguing for new artifacts to make accounting work better in practice."

regarding such issues as probable and/or pending municipal bond releases. We do not assume that users are tweeting directly about bonds; what we assume is that tweets may indirectly provide valuable insights about financial numbers of these entities. Twitter data has been found to be relevant for predictive sentiment analysis (Pak and Paroubek, 2010) but has been scarcely studied for knowledge representation.[3] During the time period of this study, PANYNJ bond series were rated as stable Aa3 by Moody's and the Port Authority Board approved the application to raise funds to upgrade one of its airports (Airport-technology news.com, 2018). It is anticipated that these announcements would evoke public reaction and an increase in public engagement, and this study applies a formal accounting ontology to the processing of these social media extractions to facilitate and organize understanding. This process is anticipated to be of interest to bond issuers, regulators, analysts, investors, auditors, and academics. This process may also be extended towards other financial domains that relate to FIBO ontologies.

## 2. Problem identification and motivation: literature review

### *2.1. Accounting and ontologies*

Accounting and ontologies have been debated in academic research extensively. This study builds on this previous work about accounting and Resources-Events-Agents (REA) ontologies. We first address what is meant by ontology and why this is important to develop a better understanding (knowledge representation/semantic meaning). Geerts and McCarthy (1999) exemplify the evolution of REA. Their paper takes an object oriented and semantic approach to REA. The paper contrasts traditional accounting systems designs with REA system designs. As a research extension to this work, the authors note the ontological directions of REA. Here, the authors contend that REA must extend its ontological features to include enterprise knowledge management, supra-accounting theories in strategic management, and an explicit treatment of time. Additionally, Mattessich (2003) gives a brief overview of the history of ontological discussions beginning with the ancient Greeks. What is important for our discussion is that in order to map physical reality onto a conceptual framework or model one must first understand what this reality consists of. Mattessich's "onion" model of reality is comprised of four layers. Mattessich (2003) applies this model to the conceptual realm of accounting. He discusses this layering in the simple example using an apple tree and the property claim to the crop of the apple tree. The apples are an observable and touchable physical reality. But if an owner has a property claim, then the income from the sale of the apples is revenue. The conditions of the apples relate to the revenue of the owner in that if the apples are rotting or bug infested, the owner's revenue will be impacted. It is the property claim which transforms the crop into income for the owner. This illustrates the hierarchical and layering nature of social reality as argued by Mattessich (2003). There are many layers of information that contribute to one's knowledge about revenue, and ontologies assist in appropriately classifying these descriptions in a standardized fashion, according to their referential meaning.

Guan et al. (2006) propose adding Bunge-Wand-Weber modeling constructs to the REA ontological model in order to solve certain deficiencies. The authors see the ontological evaluation of REA as an issue in conceptual modeling. According to the authors, conceptual modeling has two components. The first is identifying relevant phenomena in a domain. The second is mapping those phenomena into modeling constructs. According to the authors, REA has successfully implemented the first component but not the second. The authors then propose using the Bunge-Wand-Weber approach to correct these deficiencies.

Aparaschivei (2007) discusses accounting ontology from a knowledge modeling point of view. He is concerned with knowledge in accounting and artificial intelligence. He looks at the traditional hierarchy of data, information and knowledge in the accounting domain. While data and information are defined in the usual ways, that is, information is data that is useful to decision makers, he sees knowledge as being broader, deeper and richer than either data or information. In accounting he defines two types of knowledge. The first is factual domain knowledge and the second is problem solving knowledge. The first category includes law, standards, codifications, etc. The second category is partially represented in manuals but mainly is obtained from human experience and by experts. He concludes that describing an accounting ontology provides the following advantages: The first is setting up the vocabulary which will include the terms that can be used by all. The second is that the ontology facilitates interoperability between disparate systems among organizations. The third is that structuring the knowledge domain allows the representation of knowledge acquisition. And the last advantage enables knowledge sharing and reuse.

Lee (2009) expands the current discourse by looking at ontological representations of social reality. Lee is looking at the accounting context of the FASB expanding its conceptual framework to include principles-based accounting standards. He believes that such a move will need to explicitly formulate concepts of social reality into the conceptual framework. Lee, relying on Searle (1995), sees social reality as methodologically akin to linguistics. Searle describes social reality from two different perspectives, the ontological and the epistemological. Searle argues that social reality ontologically is subjective. It is created by humans subjectively through observations and consensus with regard to the function and meaning of social constructs. Further, he believes that social reality is subjective or objective depending on the truthfulness of statements made about it. This is the epistemological dimension. Lee concludes that FASB's revision of the conceptual framework is unconvincing. He believes it is a mere repackaging of the original framework. He would make a distinction between such concepts as profit or capital as social reality whereas the account classifications are primary accounting concepts.

[3] "Knowledge-representation is a field of artificial intelligence that focuses on designing computer representations that capture information about the world that can be used to solve complex problems… A knowledge representation (KR) is most fundamentally a surrogate, a substitute for the thing itself, used to enable an entity to determine consequences by thinking rather than acting, i.e., by reasoning about the world rather than taking action in it." https://en.wikipedia.org/wiki/Knowledge_representation_and_reasoning.

Lupasc et al. (2010) look at the REA framework as an ontology of accounting information systems. According to the authors, REA primitives are resources, events, agents, stock flows, control and duality. The paper treats the REA model as an ontological representation of accounting. The duality primitive is the give-and-take relationship originally mentioned in McCarthy (1982). To this the authors add a value chain concept. The authors note the extension of REA to include location. They also note that economic claims can be included in this ontology. Finally, they add the concept of an economic contract with a resulting agreement and economic commitments to the ontological model. The authors characterize these extensions as adding knowledge reuse and knowledge sharing to the ontological representation of the REA framework.

Gailly et al. use the Unified Modeling Language (UML) profile to graphically represent REA ontologies. The authors look at ontology engineering methodologies to evaluate the development of the REA ontology. The authors argue that REA is a business domain ontology. The authors conclude that REA is a specialization of a top-level ontology. Further, they determine that REA is a business domain ontology which has a universe of discourse in business. However, REA does not support all business-related subjects such as marketing strategies. According to the authors, business domain ontologies are high level ontologies. In order to be implemented, an application ontology or ontologies must be created. Therefore, they conclude that a conceptual modeling language like UML will possess the richness needed to represent REA ontological components. As such, they conclude that the REA ontology would have a semantically rich internal structure. However, the details of the structure are not explicitly specified. They then work out an explicit specification in Web Ontology Language (OWL) for REA. Ontology based research in accounting has continued and accelerated. The *Journal of Information Systems* included a special section on ontologies in its Summer 2016 issue. In 2017, the same journal published Murthy and Geets (2017) which used an REA ontology model to map big data concepts onto elements of an accounting information system. The current paper is an extension of the work in REA ontologies. The FIBO ontology has used REA as the basis for the semantic meaning of economic commitment and exchange. This paper uses the bonds part of the FIBO ontology to extract semantic meaning on bonds from Twitter streams originating from PANYNJ and MTA hashtags.

### 2.2. Twitter

Twitter is an online news and social networking service where users post and interact with SMS-like posts called “tweets”. Tweets are publicly visible by default or can be restricted by the poster to only be sent to his/her “followers” (Stutzman, 2007). Twitter is regarded as an unfiltered source of current information and news (Syed et al., 2013). As of 2016, Twitter had 319 million active users. Most of the “tweets” are of news and social networking value (Syed et al., 2013).

Twitter data is usually classified as unstructured big data (Warren et al., 2015). In fact, about 90% of big data is unstructured and is comprised of emails, social media posts such as Facebook and Twitter, phone calls, audio files and video streams (Syed et al., 2013). About 80% of the data generated by an organization is in text format, such as Word documents, memos, reports, and emails. Unstructured textual data has shown to be of increasing importance to firms that want to differentiate themselves from their competitors in areas such as market prediction, customer sentiment, and economic trends (Warren et al., 2015). Because of the volume of data, ideally, automated text mining techniques should be applied to Twitter texts to extract high-quality information. Text mining is the process of extending data mining techniques to textual information and involves finding informative models, trends, patterns, and sentiment from text files, HTML files, chats, blogs, and emails (He et al., 2013). This text mining process begins with structuring the tweets by parsing, possibly adding and/or removing certain linguistic features, and adding them to a database. The resulting structured texts are then analyzed for patterns and interpreted using text categorization, text clustering, extractions, sentiment analysis, summarization, and modeling. However, due to restrictions on the maximum number of characters (now 280 as of November 7, 2017), many tweets contain shortcuts and abbreviations which may challenge this analysis process.

Much of the research involving text mining of social media sources such as Twitter focuses on sentiment or opinion mining (Gandhi et al., 2019; Greco and Polli, 2020; Saura and Bennett, 2019; Liu and Zhang, 2012). Since the bulk of expressions found in social media comprise of individual experiences and opinions, text mining efforts have focused on the understanding and categorization of these emotions. The understanding of these opinions may enable business management to gain insights to inform better strategic decisions (Saura and Bennett, 2019). For instance, He, Zha, & Li, (2013) study the feasibility of using text mining techniques to increase competitive advantages of businesses. They analyze and monitor not only businesses' social media content, but also businesses competitors' content by compiling the sentiment of the tweets and Facebook posts regarding such topics as ordering, deliveries. Product quality, and other areas of competitive interest. They conduct a case study using text mining to analyze unstructured content from Facebook and Twitter of three large pizza chains: Domino's Pizza, Pizza Hut and Papa John's. Findings in this study suggest that social media text mining analysis could contribute greatly to gaining more insights for better organization-level decision making.

In research on consumer brand sentiment in Twitter, Mostafa (2013) analyzes 3516 tweets to evaluate polarity of consumer sentiments of brands such as IBM, T-Mobile, KLM, Nokia and DHL. He uses a predefined lexicon including around 6800 seed objectives in his text analysis. The findings show a general consumer positive sentiment towards many well-known brands. He suggests that the study contributes to the literature on consumers' general sentiment over international brands.

In stock market movement prediction, Bollen et al., 2011 conduct a sentiment text analysis of Twitter streams which showed that Twitter feeds could provide useful sentiment information or mood related to the stock market. Those moods could improve the accuracy of stock movement prediction. They investigate the public sentiment of Twitter and relate the data streams to movement in the Dow Jones Industrial Average (DJIA) by extracting a time series of the DJIA daily closing values. They find a significant correlation between moods and the DJIA closing values.

There are accepted structured lexical methods for sentiment analysis such as the Loughran-McDonald Financial Sentiment Word List (Gandhi et al., 2019; Loughran and McDonald, 2019; Loughran and McDonald, 2016). However, the process of "structuring" unstructured data or tweets to obtain high quality non-sentimental information about business and more specifically accounting and financial topics is challenging as this type of big data is unfamiliar to the accounting profession. Historically, accounting and finance professionals typically analyze numbers and have only recently expanded into textual analysis of financial statement footnotes and text, in addition to management conference calls (Warren et al., 2015). To provide structure to the textual data describing concepts of budgeting, accounting, and finance, formalization of taxonomies and hierarchies must be expanded (Moffitt and Vasarhelyi, 2013). Standardized semantic understanding and natural language processing is required to differentiate words and phrases.

### *2.3. Ontologies, text mining, and the financial domain*

Text mining and natural language processing techniques extract the underlying information, patterns, and trends from text documents (Kumar and Ravi, 2016). Many methods such as Bag of Words, N-Gram Model, Word Sense Disambiguation (WSD), Feature Selection, and Ontologies are used for text understanding and analysis. Ontologies contain domain knowledge in the form of relationships between different entities (He et al., 2013). This relationship is presented as concepts and their level order, expressed as co-occurrences and associations (Feldman and Hirsh, 1996). Kumar and Ravi (2016) observe that sentence subjectivity classification and semantic structure identification remain the most challenging problems in text mining. They acknowledge that ontologies have proven most helpful to date with these tasks and should be employed in future research in the financial domain. Additionally, they call for the development of ontologies in every domain such as finance, insurance, banking, securities, and so on (Kumar and Ravi, 2016, p. 144).

Wang et al. (2008) develop and apply ontologies to the stock trading activity of China Petroleum Corporation for stock market prediction. They present an ontology framework based on financial news articles, market investment reports, and financial analysis reports. Mellouli et al. (2010) access financial news headlines for knowledge representation using an ontology developed with the On-To-Knowledge process.[4] Cecchini et al. (2010) present a model to analyze financial circumstances with a text-based created ontology called Management Discussion and Analysis (MD & A) based on company bankruptcy filings. Their method automates ontology creation and involves steps such as preprocessing, obtaining term frequency counts, using concept frequency counts, development of the ontology using the concepts with the highest discriminatory score, scoring multi-word phrases, and finally converting the top scoring terms and concepts to a vector of values to use in subsequent analyses such as prediction and regression. Wang et al. (2011) propose a framework utilizing the ontology from Wang et al. (2008), an expert rules-based system, Bayesian Network models, and Bayesian algorithms to find the relationship between news articles and reported financial liabilities. All of these studies apply ontologies to more formalized textual sources such as reports, news headlines and articles, and sections of financial reports. What is indicated (Kumar and Ravi, 2016), is that as more textual data is being generated daily from myriad unstructured sources, more sophisticated information extraction techniques should be deployed for knowledge distillation.

### *2.4. Ontologies and social media for sentiment analysis*

One of these myriad sources of textual information is the social media platform of Twitter, which has been primarily accessed for sentiment analysis even when utilizing ontologies (Kontopoulos et al., 2013). Kontopoulos et al. (2013) propose the deployment of ontology-based techniques to Twitter for sentiment analysis. Cotfas et al. (2015) propose a new ontology-driven approach for sentiment analysis of complex feelings such as happiness, affection, surprise, anger, or sadness. This framework is not applied to financial information text. Sánchez Rada et al. (2014) propose a linked data approach for modeling sentiment and emotions in the financial domain using initiatives such as FIBO. Here the authors are concerned only with the application of a linked data approach, based on FIBO terms, to sentiments and opinions about financial concepts. The authors opine here that the sentiments provided by Twitter are important to analysts and they overlook any knowledge representation about FIBO concepts that may be available from these feeds. Our paper addresses this gap; that is, how Twitter feeds, via a FIBO ontology-based framework, may be mined for knowledge extraction in the financial domain.

## 3. Define the objectives of the solution

### *3.1. Ontology based accounting research applied to Twitter*

Although previous research discusses data standards for analysis of the softer qualitative data in financial statements (Warren et al., 2015), research has not been found that discusses formalizing financial textual information found in social media sources such as Twitter. More broadly, as discussed earlier, we found no articles in the published literature which use financial ontologies to classify social media for knowledge representation. This paper applies a frame and slot methodology from the artificial intelligence and natural language processing literature to operationalize an artifact which uses the FIBO ontology to classify Tweets in a

[4] On-To-Knowledge ontology development process comprises of two broad phases, knowledge meta and Knowledge process and follows general DSR principles within the two phases, although not identified as such.

public sector/municipalities business context for information extraction. FIBO is part of the Enterprise Data Management (EDM) Council and Object Management Group (OMG) family of specifications. FIBO provides standards for defining the facts, terms, and relationships associated with financial concepts. FIBO concepts are vetted by subject matter experts (SMEs) so they should reflect high quality financial concepts. We show that this grammar can be used to mine knowledge representation from unstructured textual data. We compare this to an approach using naïve key words and then use a false positive metric to score the artifact. Twitter streams were monitored and analyzed with an automated code based on this FIBO derived framework. With such analysis, constituent semantic structures were uncovered which expose reactions to policies and programs more quickly than by following the feeds manually.

## 4. Design and development of an artifact which meets some of the objectives

### *4.1. Derivation of the slot and frame structure*

FIBO is a very rich knowledge representation of the financial industry. The content teams consist of subject matter experts (SMEs) in the various subareas of finance such as business entities and derivatives. This research uses only a small subset of the knowledge available in FIBO, the representation of municipal bonds. This research uses the representation of financial bonds for illustrative purposes only; it is expected that our artifact could be applied to any other FIBO concept or domain. While it is possible to use FIBO as a collection of terms with which to get key words for analyzing a twitter stream, doing so would lose the knowledge base aspect of the FIBO ontology. As such, frame representations are used here to maintain the knowledge imbedded in the FIBO ontology in order to help filter tweets to find threads referencing municipal bonds. Furthermore, this research uses the synonyms and broader synonyms also available in the FIBO specification to further enrich the semantic classification ability of our artifact. We further explain the artifact in the text that follows.

Frames are useful for representing knowledge. Frames are used to represent related knowledge about narrow subjects which have a lot of knowledge context. Frames are good choices to represent physical or conceptual structures, such as cars and financial instruments. A frame can be thought of as similar to a record structure, where the fields and field values of a record are the slots and slot fillers of a frame. A frame is basically a set of slots that define an object (Table 1).

Frames are designed generally to represent either generic or specific knowledge. Slot fillers can also contain relations such as the is-a or a-kind-of relations. So, a generic car frame may be represented as in Table 2.

Frame hierarchy systems are designed so that more generic frames are at the top of the hierarchy. Frames model real-world objects by using generic knowledge for most of the object's attributes while using specific knowledge for special cases. An object which has all the typical attributes is called a prototype. Frames are also classified by their applications. A situational frame contains knowledge about what to expect in a specific situation. An action frame contains slots that specify actions that are performed in specific situations. A combination of situational and action frames can be used to represent cause-and-effect relationships in what are known as causal knowledge frames (Table 3).

Since Twitter feeds tend to contain minimal detailed information, they are unlikely to conform to highly complex frame structures. This being the case, we reduce the government issued bonds portion of FIBO into a frame structure using concepts as the slots except in the case of the lowest level concepts where these become slot fillers. A portion of the FIBO diagram for municipal bonds is shown in Fig. 1.

The resulting frame for government issued bonds is then as follows (Table 4):

Some of these FIBO concepts also contain synonyms and near synonyms as specified in FIBO. We use these terms as well when classifying Tweets.

### *4.2. Design of the solution*

#### *4.2.1. Design of the generic ontology-based twitter knowledge extraction model*

Based on the approaches of previous Twitter studies and of accounting ontology applications (Du and Zhou, 2012), and after referencing the appropriate ontology (FIBO) and conceptualizing the appropriate frames and their slots and synonyms, the components of our proposed generic knowledge extraction artifact are as follows:

1. *Target a population of Tweets* - Decide which Twitter streams are relevant and identify their critical Twitter keys, and decide on a time frame
2. *Develop an Application Programming Interface (API) for the targeted Tweets* – Decide how to access the Twitter feeds from https://twitter.com/.
3. *Collect and assemble the data* – Write a program code[5] to access the Twitter API to fetch all Twitter streams that contain at least one of the targeted keys from Step 1.
4. *Data aggregation and pre-processing* – We aggregate six fields from each tweet: date/time, tweet content, user ID, number of followers, number of likes, number of posts. We pre-process the data by removing commas and then uploading the data into a database.

[5] We coded our extraction using Python 2.7 (https://www.python.org/). The code is available upon request from the corresponding author.

**Table 1**
A typical frame with frame object "car".

| Slot | Filler |
|---|---|
| Maker | Ford |
| Model | F-150 |
| Year | 2013 |
| Engine | Gasoline |
| Tires | Goodyear |
| Color | Blue |

**Table 2**
Generic car frame.

| Slot | Filler |
|---|---|
| Name | Car |
| Specialization-of | A-kind-of property |
| Types | (SUV, compact, luxury) |
| Maker | (Honda, Ford, Subaru) |
| Engine | (Gasoline, hybrid, diesel, electric) |
| Transmission | (Manual, automatic) |

5. *FIBO term search* – Search the databases for the data structures that fill the slots of the frame of the ontology
6. *Construct validity test (ontology- versus naive keywords)* – To check for validity of the ontology application, a comparison to a naïve (non-ontology expert) keyword search should be conducted to see if the ontology performance is superior.
7. *Design of the descriptive statistics algorithm and generic classifiers* – Collect all the matches to the ontology frame, including synonyms and near-synonyms.
8. *Tuning the FIBO generic classifiers* – Test all possible combinations of classifiers identified in step 7 to see if any combinations of several frames perform better than the stand-alone classifiers to create a "tuned classifier". For Tweets, this tuning process is

**Table 3**
An instance of a car frame.

| Slot | Filler |
|---|---|
| Name | Bob's car |
| Specialization-of | Is a car |
| Type | Luxury |
| Maker | Subaru |
| Engine | Hybrid |
| Transmission | Automatic |

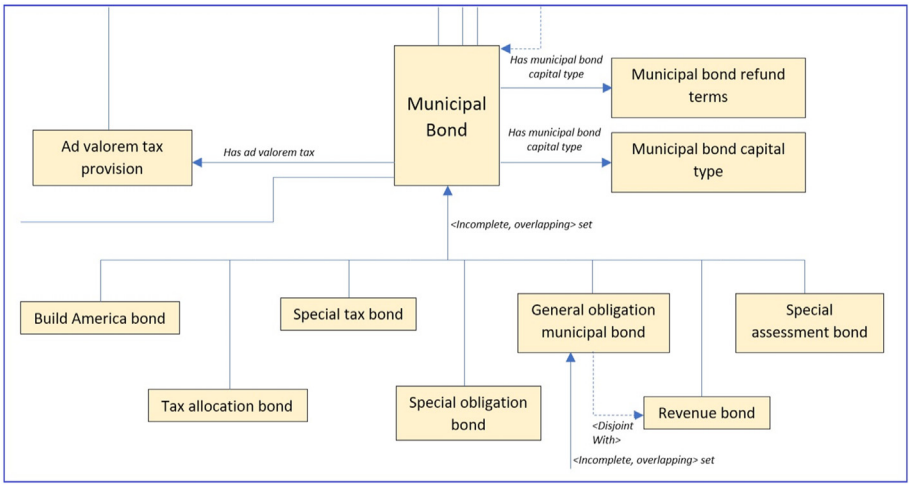


**Fig. 1.** FIBO municipal bonds.

**Table 4**
Government Issued Bond Frame from FIBO.

| Slot | Filler |
|---|---|
| Government issued bond | |
| Municipal security | |
| Municipal debt issuer | |
| Municipal bond | |
| Debt obligor | |
| Funds usage | |
| Municipal bond capital type | |
| Municipal bond refund terms | |
| Municipal trustee | |
| Ad valorem tax provision | |
| Municipal bond type | (Build America, Tax allocation, special tax, special obligation, general obligation, revenue, special assessment, consolidated bond) |

not expected to improve results as Tweets are generally too short. However, for non-Tweet data sources, this step could be informative and therefore is included here.

9. *Using the tuned classifiers to classify other twitter feeds* – Once an optimal "tuned classifier" is arrived at, it should be applied to other collected Tweets to possibly improve results. This could be applied to an untrained/untested Twitter dataset or to a Twitter stream that was not used to develop the "tuned classifier".
10. *Evaluation of the final results* – Evaluation of the results of the classifiers in terms of dominant themes, time frames, and other descriptive statistics. The results of the "tuned classifiers" are evaluated for their ability to facilitate an accurate extraction of any Twitter data that is relevant to the project problem.

The sequential workflows of this generic extraction artifact are outlined in Fig. 2.

Each of the steps in the artifact is described more fully here, customized for our application of knowledge extraction from Twitter about municipal bonds by use of the FIBO ontology:

1. Targeted Tweets: The first step of our framework. Initially, we have chosen two major transportation agencies: The Port Authority of New York and New Jersey (PANYNJ), and The Metropolitan Transportation Authority (MTA). Also, one major division of PANYNJ is the PATH subway system. One reason for choosing PANYNJ and the MTA systems is because they are responsible for nearly all of New York and northern New Jersey's transportation infrastructure – subways, buses, commuter rail, bridges, tunnels, airports, and ports. In order to expand our search, we add as many words (keys) as possible, targeting all possible public tweets that mention those transportation hubs. For instance, to fetch tweets mentioning PANYNJ, we include the following keys (PANYNJ, PORTAUTHORITY, PORT AUTHORITY, PABusTerminal, PORTNYNJ, Port Authority NY&NJ, the Port authority, Port authority of NY & NJ, Port authority of NY and NJ, the Port of New York and New Jersey). For tweets that mention MTA, the following keys were included (NYCT Subway, #MTA, #MTATransparency, NYCTSubway, NYCTBus, @MTA, LIRR, NYC Subway, #nycsubway). Finally, for tweets mentioning the PATH, the following keys are included (Path train, Path service, Pathtrain, #pathtrain).
2. Twitter API: The second step of our framework. After deciding on what will be the focus and scope of the study, the Twitter Micro-blogging social media platform is chosen as the source of the data (https://twitter.com/). Twitter enables millions of

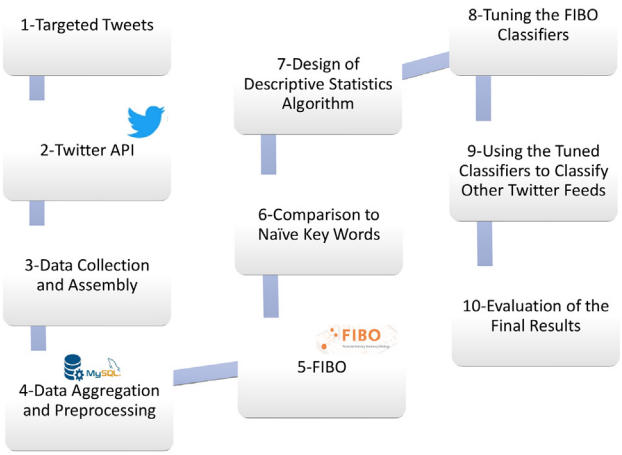


**Fig. 2.** Generic ontology (FIBO)-Twitter knowledge extraction.

**Fig. 3.** PANYNJ Twitter live-stream (Python 2.7).

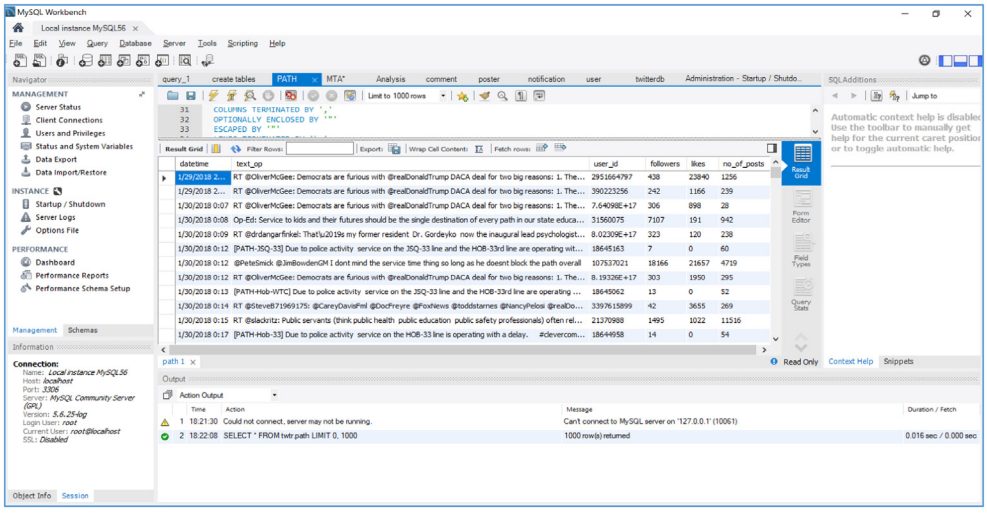

**Fig. 4.** The PATH table stored in MySQL workbench.

users to share their feelings, opinions regarding any issue, any day. As a result, Twitter is considered as a rich source of data for opinion mining and sentiment analysis, Pak & Paroubek (2010).

3. Data Collection and Assembly: In this step, by accessing the Twitter Application Programming Interface (API), we write a Python code using Python 2.7 (https://www.python.org/) to fetch all Twitter streams that contains at least one of the targeted keys mentioned in Step One. Many Python libraries are used such as Tweepy, which allows us to access the Twitter API and StreamListener, a method that allows us to stream real-time messages (tweets) and route them to a storage space. We run the code at the same time to stream real-time data, one for each agency (PANYNJ, PATH, MTA). The data is collected from January 29th until October 19th, 2018. The following Fig. 3 is a screenshot of the live-stream from Twitter API running in the background:

4. Data Aggregation and Preprocessing: After collecting the raw data, we aggregate six fields from each tweet. The fields are the following: date and time of tweet, original message (tweet), user identification number, number of followers, number of likes, number of posts that a certain user has. Then, we perform a preprocessing step, first by removing the commas from the text body, then uploading the data into a database. We use a Structured Query Language Platform, MySQL. MySQL is an open-

**Table 5**
The PANYNJ, MTA, PATH tables.

| Table name | # of tweets | Aggregation period |
|---|---|---|
| PANYNJ | 101,634 | 1/29/2018–10/19/2018 |
| MTA | 541,542 | 1/29/2018–10/19/2018 |
| PATH | 106,222 | 1/29/2018–10/19/2018 |
| **Total** | **749,398** | 1/29/2018–10/19/2018 |

source relational database management system (https://www.mysql.com/). We load all the fetched data to MySQL using the 'Load Data' statement which reads rows from text or comma separated files (csv) into SQL tables. Fig. 4 shows a screenshot of part of the PATH table from the MySQL database management system.

After initial data aggregation and preprocessing during the period from Jan 29th, 2018 to October 19th, 2018, the intermediate datasets consist of the following (Table 5):

The total number of records means that the number of tweets or retweets that include at least one of the key terms mentioned previously for each of the agencies.

5. FIBO term search: After data collection, aggregation and preprocessing, we search the databases for data structures which fill the slots of the frame for government bonds developed from the FIBO ontology. We use the framework derived in the next section to perform the search. We search both on individual tweets and on threads.
6. Construct validity test: After collecting all the tweets related to the bond information of the two agencies, we compare the results to a naïve key word search of the database. This indicates whether the semantics provided by the FIBO SMEs is superior to using simple key word searches and is a form of model validation for our artifact. Our metric is the number of false positives[6] of each classification method. This serves as a test of the construct validity of using FIBO terms as opposed to naive keywords.
7. Design of the Descriptive Statistics Algorithm: In this step we collect the matches for all the terms in the FIBO frame, including synonyms and near synonyms, for the MTA and PANYNJ twitter streams. This allows us to see which slots and terms are relevant for this application. This step results in the creation of two generic classifiers: one derived from the MTA stream and the other from the PANYNJ stream. We then "tune" these generic classifiers to see if we could improve their performance in terms of correct classification and the false positives metric.
8. Tuning the FIBO classifiers: In this step we use each of the generic classifiers to see whether we could improve its results. We did this by testing whether classifiers built from combinations of several frames perform better than the generic classifiers. Since individual tweets are very short, we do not expect to improve performance. Our expectations end up being correct.
9. Using the tuned classifiers to classify other twitter feeds: In our final test, we use the tuned MTA classifier to classify the PANYNJ twitter stream and the tuned PANYNJ classifier to classify the MTA twitter stream. We then use each of the tuned classifiers to classify the PATH twitter stream.
10. Evaluation of Final Results: In this final step, we evaluate results of the MTA and PANYNJ classifiers. With such evaluation of final results, we provide a proof of concept for using FIBO frames in text mining as constituent semantic structures.

Steps 1 through 5 above are followed to produce the three twitter streams (MTA, PANYNJ and PATH). The remaining part of the paper discusses steps 6 through 10 in additional detail and then concludes with parting thoughts and suggestions for future research.

## 5. Demonstration of the solution

### *5.1. Initial testing: construct validity test*

The following tables show an illustration of the findings between FIBO terms search and the naïve terms search for the three twitter streams: MTA, PANYNJ and PATH. The results only show the application of FIBO concepts, not the synonyms and near synonyms. Adding the latter would only improve the results. False positives are tweets which were initially classified as being relevant but are not. An example would be if someone tweets on the MTA feed about the muddy "path" by the Hudson River.

**Table 6**
MTA FIBO terms search.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP |
|---|---|---|---|---|
| 280 | 138 | 125 | 20 | 16% |

**Table 7**
MTA naïve terms search.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP | Ratio of # of tweets of naïve to FIBO |
|---|---|---|---|---|---|
| 85 | 36 | 41 | 18 | 43.9% | 30.4% |

[6] We also looked at the occurrences of false negatives in our population by taking a random 1000 tweets from the MTA table. The results show the percentage of false negatives is small and equals to 26 tweets which is only 2.6%.

The Port Authority of New York and New Jersey (PANYNJ) tweets:

**Table 8**
PANYNJ FIBO terms search.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP |
|---|---|---|---|---|
| 117 | 61 | 54 | 1 | 1.9% |

**Table 9**
PANYNJ naïve terms search.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP | Ratio of # of tweets of naïve to FIBO |
|---|---|---|---|---|---|
| 31 | 7 | 22 | 4 | 18.2% | 26.5% |

The Port Authority Trans-Hudson (PATH) tweets:

**Table 10**
PATH FIBO terms search.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP |
|---|---|---|---|---|
| 48 | 14 | 34 | 21 | 61.8% |

**Table 11**
PATH naïve terms search.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP | Ratio of # of tweets of naïve to FIBO |
|---|---|---|---|---|---|
| 2 | 1 | 1 | 1 | 100% | 4.2% |

In our initial validation above (Tables 6–11) we compare the results between the FIBO terms search which is present in Table 4 (excludes the synonyms) and the naïve terms search. One can clearly notice that the FIBO terms search is returning more records compared to naïve terms search in all datasets. For instance, looking at the MTA dataset, one can notice a total of 280 records retrieved using FIBO terms search compared to 85 records using naïve terms search. Also, the percentage of false positives is significantly higher in naïve terms search compared to FIBO search with a total of 43.9% compared to 16% respectively. Similarly, looking at PANYNJ dataset we find a 26.5% total ratio of number of tweets of naïve to FIBO terms searches which is significantly lower. In addition, we also find that there is a low percentage of false positives, 1.9%, using FIBO compared to 18.2% of false positives using naïve terms search. In the PATH tables (Tables 10–11), we also notice a higher number of records retrieved using FIBO terms search compared to naïve terms search with less false positives; however, 61.8% of false positives using the FIBO search is still considered a high number but it is still lower than the naïve terms search (100% FP).

So, from completing the initial validation we can clearly see that by using FIBO terms search, we can retrieve more records compared to naïve terms search. This is an indication that the SMEs' expertise has been captured in the FIBO ontology and was successfully transferred into our artifact. Additionally, in two out of the three cases, we observe a lower percentage of false positives using the FIBO terms search. In the next section, we will further classify the tweets to see if we can obtain more accurate results by getting the synonyms and apply the frame and slots methodology. Please note that the percentage of false positives is calculated based on the unique number of tweets.

### *5.2. Demonstration of the solution continued: design of the implemented system on a real twitter feed*

In our artifact, we derive the following frame and slots terms and their synonyms from the FIBO ontology Municipal Bonds Full (Table 12):

For the concepts we add to the algorithm synonyms from the FIBO ontology that refer to broader topics, since these words may be used in natural language (where the context is implicit) to refer to these concepts. These are synonyms and near synonyms in the FIBO specification. Since the FIBO terms, synonyms and near synonyms are all developed by subject matter experts in the financial industry, we expect that searches which include a broader selection of terms will find more tweets which discuss

**Table 12**
Frame and slots terms and their synonyms from the FIBO ontology Municipal Bonds Full (FIBO Frame).

| Concept | Synonym | Near/broader synonym or role-performing item |
|---|---|---|
| Government issued bond | Sovereign bond, treasury bond | Government debt |
| Municipal security | Municipal debt instrument | Muni |
| Municipal debt issuer | Muni issuer | Issuer, municipality |
| Municipal bond | Muni bond, Muni | |
| Debt obligor | Owing party, borrower | Obligor |
| Funds usage | Funds purpose, disbursement purpose | Loan purpose, credit facility purpose, credit purpose |
| Municipal bond capital type | Muni capital type | Capital type |
| Municipal bond refund terms | Muni refund terms | Refund terms |
| Municipal trustee | Muni trustee | Trustee |
| Ad valorem tax provision | | Property tax provision, real property tax provision, sales tax provision |
| Municipal bond type | N/A | |
| Build America | Build America bond | |
| Tax allocation | Tax allocation bond | |
| Special tax | Special tax bond | |
| Special obligation | Special obligation bond | |
| General obligation | General obligation bond | |
| Revenue | Revenue bond | |
| Special assessment | Special assessment bond | |

municipal bonds. We anticipate that very few tweets would discuss municipal bonds explicitly and that the public will tweet more often about topics relating to bonds and their FIBO concepts and synonyms/near synonyms.

### 5.3. Construction of the frame-based system

A frame consists of set of slots which are filled by values, procedures, or links to other frames. We need to formalize the municipal bonds-type frames taken from FIBO as shown in Table 12.

#### 5.3.1. Design of the descriptive statistics algorithm

For our design artifact, our algorithm performs multiple tests aiming at finding what best constitutes a frame, that is, to tune the results given which slots and synonyms are best at classifying the Twitter feed. The frame and slots terms and their synonyms and near synonyms are taken from Table 12, which is derived from the FIBO ontology Municipal Bonds Full. We assume the slots are represented as they appear in Table 13 (note the partial representation of Table 12 for illustration purposes on how to pick the slots):

**Table 13**
Illustration of the slots of the frame and the synonyms from FIBO ontology Municipal Bonds Full.

| Concept | Synonym | Near/broader synonym or role-performing item |
|---|---|---|
| $S_{1,1}$ = Government issued bond | $S_{1,2}$ = Sovereign bond, treasury bond | $S_{1,3}$ = Government debt |
| $S_{2,1}$ = Municipal security | $S_{2,2}$ = Municipal debt instrument | $S_{2,3}$ = Muni |
| $S_{3,1}$ = Municipal debt issuer | $S_{3,2}$ = Muni issuer | $S_{3,3}$ = Issuer, municipality |
| … | … | … |
| $S_{18,1}$ = Special assessment | $S_{18,2}$ = Special assessment bond | |

Definition of the Text Terms:
The set of text terms would follow similar logic but by adding additional dimension to include terms from Column 1 + Column 2 + Column 3 of Table 12 as follows:
$T = S_{1,1} \vee S_{1,2} \vee S_{1,3} \vee S_{2,1} \vee S_{2,2} \vee S_{2,3} \vee S_{3,1} \vee S_{3,2} \vee S_{3,3} \vee S_{4,1} \vee S_{4,2} \vee S_{4,3} \ldots S_{18,1} \vee S_{18,2} \vee S_{18,3}$.
SQL query:

```
SELECT * FROM TWTR.(PANYNJ, PATH, MTA) WHERE text_op LIKE '% S1,1 %
OR text_op LIKE '% S1,2 %'
OR text_op LIKE '% S1,3 %'
OR text_op LIKE '% S2,1 %'
```

OR text_op LIKE '% $S_{2,2}$ %'
OR text_op LIKE '% $S_{2,3}$ %'
OR text_op LIKE '% $S_{3,1}$ %'
OR text_op LIKE '% $S_{3,2}$ %'
OR text_op LIKE '% $S_{3,3}$ %'
OR text_op LIKE '% $S_{4,1}$ %'
OR text_op LIKE '% $S_{4,2}$ %'
OR text_op LIKE '% $S_{4,3}$ %'
OR …
OR text_op LIKE '% $S_{18,1}$ %'
OR text_op LIKE '% $S_{18,2}$ %';
OR text_op LIKE '% $S_{18,3}$ %'

The definitions above are part of our design artifact for frame construction. In these definitions, we use the logical condition 'OR' or '∨' for the slots. So, the question is what constitutes or best constitutes a frame? In order to explore whether all slots need to be included in the search or not, we develop descriptive statistics about the frequency of occurrence of each term in Table 12. We also develop statistics on the frequency of the slots being filled with retweets excluded. Based on these descriptive statistics, we proceed by developing an efficient method for extracting frames from the Twitter stream. Efficiency here is defined by how many slots we need to best represent a frame.

*5.3.2. Descriptive statistics*

In this section, we report on the descriptive statistics of our population. We use two of our three tables which are the MTA and PANYNJ. Tables 14, 15, and 16 show the FIBO terms search for the system on our population (real Twitter feed) for The MTA and The PANYNJ datasets:

After our algorithm provides the descriptive statistics on the MTA and PANYNJ datasets, we find that of the eighteen FIBO slots we have, only six slots are filled by at least one value. From this we code the program to construct generic classifiers which only use FIBO terms for which we found results in the descriptive analysis. From the results in Tables 14 and 15, the two generic classifiers are:

**Table 14**
Frequency of the frames and slots terms and their synonyms from the FIBO ontology Municipal Bonds Full for the MTA dataset.

| FIBO concepts | # of tweets | FIBO synonyms | # of tweets | Contextualized synonym or role-performing item | # of tweets | Frequency of tweets by row |
|---|---|---|---|---|---|---|
| Government issued bond | 0 | Sovereign bond | 0 | Government debt | 0 | 0 |
| | | Treasury bond | 0 | | | |
| Municipal security | 0 | Municipal debt instrument | 0 | N/A | N/A | 0 |
| Municipal debt issuer | 0 | Muni issuer | 0 | Issuer | 0 | 0 |
| Municipal debt | 1 | | | Municipality | 3 | 0 |
| Municipal bond | 0 | Muni bond | 3 | N/A | N/A | 0 |
| | | Muni | 10 | | | 0 |
| Debt obligor | 0 | Owing party | 0 | Obligor | 0 | 0 |
| | | Borrower | 0 | | | 0 |
| Funds usage | 0 | Funds purpose | 0 | Loan purpose | 0 | 0 |
| | | | | Credit facility purpose | 0 | 0 |
| | | Disbursement purpose | 0 | Credit purpose | 0 | 0 |
| Municipal bond capital type | 0 | Muni capital type | 0 | Capital type | 0 | 0 |
| Municipal bond refund terms | 0 | Muni refund terms | 0 | Refund terms | 0 | 0 |
| Municipal trustee | 0 | Muni trustee | 0 | Trustee | 5 | 0 |
| Ad valorem tax provision | 0 | N/A | N/A | Property tax provision | 144 | 0 |
| | | | | Real property tax provision | 0 | 0 |
| | | | | Sales tax provision | 21 | 0 |
| Municipal bond type | 0 | N/A | N/A | N/A | N/A | |
| Build America | 0 | Build America bond | 0 | N/A | N/A | 0 |
| Tax allocation | 0 | Tax allocation bond | 0 | N/A | N/A | 0 |
| Special tax | 1 | Special tax bond | 0 | N/A | N/A | 0 |
| Special obligation | 0 | Special obligation bond | 0 | N/A | N/A | 0 |
| General obligation | 0 | General obligation bond | 0 | N/A | N/A | 0 |
| Revenue | 253 | Revenue bond | 2 | N/A | N/A | 2 |
| Special assessment | 0 | Special assessment bond | 0 | N/A | N/A | 0 |

**Table 15**
Frequency of the frames and slots terms and their synonyms from the FIBO ontology Municipal Bonds Full for the PANYNJ dataset.

| FIBO concepts | # of tweets | FIBO synonyms | # of tweets | Contextualized synonym or role-performing item | # of tweets | Frequency of tweets by row |
|---|---|---|---|---|---|---|
| Government issued bond | 0 | Sovereign bond | 0 | Government debt | 0 | 0 |
| | | Treasury bond | 0 | | | |
| Municipal security | 0 | Municipal debt instrument | 0 | N/A | N/A | 0 |
| Municipal debt issuer | 0 | Muni issuer | 0 | Issuer | 0 | 0 |
| Municipal debt | 0 | | | Municipality | 8 | 0 |
| Municipal bond | 2 | Muni bond | 0 | N/A | N/A | 0 |
| | | Muni | 0 | | | |
| Debt obligor | 0 | Owing party | 0 | Obligor | 0 | 0 |
| | | Borrower | 0 | | | 0 |
| Funds usage | 0 | Funds purpose | 0 | Loan purpose | 0 | 0 |
| | | | | Credit facility purpose | 0 | 0 |
| | | Disbursement purpose | 0 | Credit purpose | 0 | 0 |
| Municipal bond capital type | 0 | Muni capital type | 0 | Capital type | 0 | 0 |
| Municipal bond refund terms | 0 | Muni refund terms | 0 | Refund terms | 0 | 0 |
| Municipal trustee | 0 | Muni trustee | 0 | Trustee | 8 | 0 |
| Ad valorem tax provision | 0 | N/A | N/A | Property tax provision | 18 | 0 |
| | | | | Real property tax provision | 0 | 0 |
| | | | | Sales tax provision | 0 | 0 |
| Municipal bond type | 2 | N/A | N/A | N/A | N/A | 2 |
| Build America | 0 | Build America bond | 0 | N/A | N/A | 0 |
| Tax allocation | 0 | Tax allocation bond | 0 | N/A | N/A | 0 |
| Special tax | 3 | Special tax bond | 0 | N/A | N/A | 0 |
| Special obligation | 0 | Special obligation bond | 0 | N/A | N/A | 0 |
| General obligation | 0 | General obligation bond | 0 | N/A | N/A | 0 |
| Revenue | 111 | Revenue bond | 7 | N/A | N/A | 7 |
| Special assessment | 0 | Special assessment bond | 0 | N/A | N/A | 0 |

**Table 16**
Frequency of the frames and slots terms and their synonyms from the FIBO ontology Municipal Bonds Full for the PATH dataset.

| FIBO concepts | # of tweets | FIBO synonyms | # of tweets | Contextualized synonym or role-performing item | # of tweets | Frequency of tweets by row |
|---|---|---|---|---|---|---|
| Government issued bond | 0 | Sovereign bond | 0 | Government debt | 2 | 0 |
| | | Treasury bond | 0 | | | 0 |
| Municipal security | 0 | Municipal debt instrument | 0 | N/A | N/A | 0 |
| Municipal debt issuer | 0 | Muni issuer | 0 | Issuer | 0 | 0 |
| Municipal debt | 0 | | | Municipality | 2 | 0 |
| Municipal bond | 0 | Muni bond | 0 | N/A | N/A | 0 |
| | | Muni | 1 | | | 0 |
| Debt obligor | 0 | Owing party | 0 | Obligor | 0 | 0 |
| | | Borrower | 0 | | | 0 |
| Funds usage | 0 | Funds purpose | 0 | Loan purpose | 0 | 0 |
| | | | | Credit facility purpose | 0 | 0 |
| | | Disbursement purpose | 0 | Credit purpose | 0 | 0 |
| Municipal bond capital type | 0 | Muni capital type | 0 | Capital type | 0 | 0 |
| Municipal bond refund terms | 0 | Muni refund terms | 0 | Refund terms | 0 | 0 |
| Municipal trustee | 0 | Muni trustee | 0 | Trustee | 8 | 0 |
| Ad valorem tax provision | 0 | N/A | N/A | Property tax provision | 1 | 0 |
| | | | | Real property tax provision | 0 | 0 |
| | | | | Sales tax provision | 0 | 0 |
| Municipal bond type | 0 | N/A | N/A | N/A | N/A | 0 |
| Build America | 0 | Build America bond | 0 | N/A | N/A | 0 |
| Tax allocation | 0 | Tax allocation bond | 0 | N/A | N/A | |
| Special tax | 0 | Special tax bond | 0 | N/A | N/A | 0 |
| Special obligation | 0 | Special obligation bond | 0 | N/A | N/A | 0 |
| General obligation | 0 | General obligation bond | 0 | N/A | N/A | 0 |
| Revenue | 48 | Revenue bond | 0 | N/A | N/A | 0 |
| Special assessment | 0 | Special assessment bond | 0 | N/A | N/A | |

MTA: Municipal debt OR municipality OR Muni bond OR muni OR trustee OR property tax provision OR sales tax provision OR special tax OR revenue OR revenue bond; and,
PANYNJ: Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond.

Please note that we decide to discard the PATH table going forward since it returns only a few values of the frames and slots.

Our algorithm now takes the slots with more than one value and sees if we can get better results[7] (fewer false positives) when we tune the system by requiring that multiple slots are used to classify a tweet as being more specifically about bonds. We then construct automated classifiers from the descriptive statistics by testing pairwise combinations of the slots with their synonyms and near synonyms for slots having more than one match. Since an AND condition can at most only include the already tested OR conditions, we only test the pairwise comparisons. If our results had been different, we would have continued to test triples, etc. Given the twitter streams in this study, this proves unnecessary. The two sets of tuned classifiers based on the descriptive statistics above are as follows.

MTA
1. (Property Tax OR Sales Tax Provision) AND Trustee
2. (Property Tax OR Sales Tax Provision) AND Municipality
3. (Property Tax OR Sales Tax Provision) AND (Muni Bond OR Muni)
4. (Property Tax OR Sales Tax Provision) AND (Revenue OR Revenue Bond)
5. (Revenue OR Revenue Bond) AND Trustee
6. (Revenue OR Revenue Bond) AND Municipality
7. (Revenue OR Revenue Bond) AND (Muni Bond OR Muni)

PANYNJ
1. (Revenue OR Revenue Bond) AND Municipality
2. (Revenue OR Revenue Bond) AND Municipal Bond
3. (Revenue OR Revenue Bond) AND Trustee
4. (Revenue OR Revenue Bond) AND Property Tax
5. (Revenue OR Revenue Bond) AND Municipal Bond Type
6. (Revenue OR Revenue Bond) AND Special Tax.

*5.3.3. Tuning the results*

We found the following classification results during tuning:

MTA
1. (Property Tax OR Sales Tax Provision) AND Trustee = 0
2. (Property Tax OR Sales Tax Provision) AND Municipality = 0
3. (Property Tax OR Sales Tax Provision) AND (Muni Bond OR Muni) = 0
4. (Property Tax OR Sales Tax Provision) AND (Revenue OR Revenue Bond) = 12, Zero false positive. 100% true positive.
5. (Revenue OR Revenue Bond) AND Trustee = 0
6. (Revenue OR Revenue Bond) AND Municipality = 0
7. (Revenue OR Revenue Bond) AND (Muni Bond OR Muni) = 0

PANYNJ
1. (Revenue OR Revenue Bond) AND Municipality = 0
2. (Revenue OR Revenue Bond) AND Municipal Bond = 0
3. (Revenue OR Revenue Bond) AND Trustee = 0
4. (Revenue OR Revenue Bond) AND Property Tax = 0
5. (Revenue OR Revenue Bond) AND Municipal Bond Type = 0
6. (Revenue OR Revenue Bond) AND Special Tax = 0.

Therefore, we can derive the best classifiers as follows:

MTA:

In Table 17 we show Municipal debt OR municipality OR Muni bond OR muni OR trustee OR property tax provision OR sales tax provision OR special tax OR revenue OR revenue bond (the generic classifier from Table 14).

[7] The following two tweets show an example of what is relevant (True and false positives). An example of a true positive tweet taken from our population would be like: "LIBOR Switch - latest SOFR issuance via muni bond from New York MTA. Strong demand leading a first roll of US$107 m."; A false positive tweet example: "So at work today a discussion was had about the #NY Tolls #MTA property tax and the horrible job @andrewcuomo is doing."

**Table 17**
Results from the best MTA classifier.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP |
|---|---|---|---|---|
| 423 | 256 | 167 | 20 | 12% |

**Table 18**
Results from the best PANYNJ classifier.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP |
|---|---|---|---|---|
| 148 | 75 | 73 | 13 | 17.8% |

Compare this with the only other candidate for MTA: (Property Tax OR Sales Tax Provision) AND (Revenue OR Revenue Bond) = 12, FP 0%. If we choose this, #4, we lose 147 (true positives) – 12 = 135 Tweets which are false negatives for the #4 classifier. This is too much loss of information for a small gain in precision.

PANYNJ:

We demonstrate in Table 18 Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond (the generic classifier from Table 15).

This has 148 classifications with a 17.8% FP rate. There are no other candidates here, so we choose the OR classifier above.

*5.3.4. Using the tuned classifiers to classify other Twitter feeds*

In Table 19 we use the best MTA classifier to classify the whole PANYNJ population which is the following:

"Municipal debt OR municipality OR Muni bond OR muni OR trustee OR property tax provision OR sales tax provision OR special tax OR revenue OR revenue bond".

Then we use the best PANYNJ classifier to classify the whole MTA population which is the following and is shown in Table 20:

"Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond".

**Table 19**
Results of classifying PANYNJ data with the MTA best classifier.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP |
|---|---|---|---|---|
| 148 | 76 | 69 | 3 | 4.3% |

**Table 20**
Results of classifying MTA data with the best PANYNJ classifier.

| # of tweets | Retweets | Unique # of tweets | False positives | % of FP |
|---|---|---|---|---|
| 418 | 268 | 150 | 7 | 4.7% |

*5.3.5. Testing our artifact on a new population*

In this section, we test our artifact using new data collected from October 22, 2018 to March 10, 2019. Table 21 shows the number of records each table contains.

**Table 21**
Testing data - The MTA and PANYNJ tables.

| Tables | Raw data | Cleaned data |
|---|---|---|
| MTA | 323,413 | 322,110 |
| PANYNJ | 6956 | 6904 |

*5.3.6. Testing our tuned classifiers using the new dataset*

In the summary Table 22, we use the previously introduced best MTA classifier "Municipal debt OR municipality OR Muni bond OR muni OR trustee OR property tax provision OR sales tax provision OR special tax OR revenue OR revenue bond" to classify the new MTA and PANYNJ populations. Similarly, we use the best PANYNJ classifier "Municipality OR municipal bond OR Trustee OR property tax provision OR Municipal bond type OR special tax OR revenue OR revenue bond" to classify both the new MTA and PANYNJ populations.

**Table 22**
Performance assessment of various classifiers applied on our MTA and PANYNJ test data.

| Data | Classifier | # of tweets | Retweets | Unique # of tweets | False positives | Misclassification rate (% FP) | Accuracy | Recall |
|---|---|---|---|---|---|---|---|---|
| MTA | PANYNJ | 506 | 270 | 236 | 2 | 0.8% | 0.99 | 1.0 |
| MTA | MTA | 527 | 272 | 255 | 5 | 1.9% | 0.98 | 1.0 |
| PANYNJ | PANYNJ | 31 | 27 | 4 | 0 | 0 | 1.0 | 1.0 |
| PANYNJ | MTA | 31 | 27 | 4 | 0 | 0 | 1.0 | 1.0 |

The results above show that by testing our tuned classifiers (i.e., best classifiers) in a new population we get similar results, or even more accurate output. Using the best MTA classifier to classify the new MTA dataset, we see an increase in precision of 1.9% compared to 12% which is the percentage of false positives or misclassification rate. Moreover, the total number of records captured is 527 compared to 423 records. In addition, the MTA classifier has an accuracy of 98%. Similarly, by applying the best PANYNJ classifier on the new population (i.e., the new MTA dataset), we see an improvement in precision that is 0.8% compared to 4.7% of false positives and an accuracy of 99%. We can also notice the increased number of records fetched using our classifiers compared to the training phase. The total number of records went up to slightly above 500 in both classifiers compared to around 400 records even though the duration of data collection is almost half the time in the testing-data-collection period compared to the initial datasets (10-months compared to 5-months). We think this is because of the controversial issues that occur during the testing phase such as Amazon's reneging on the opening of its new HQ2 in Long Island City, NY, and how this could cause a huge loss in tax revenue that could have been used to support NYC infrastructure (e.g., MTA). Also, the MTA board approved fare and toll increases on subways, railroads, buses and tolled crossings on Feb 27, 2019.

Similarly, by running our tests on the new PANYNJ population, we notice similar findings compared to our previous tests; that is, using the tuned classifiers (i.e., best MTA and PANYNJ classifiers) to classify other twitter feeds (The PANYNJ new population). We find that we almost receive similar percentages of false positives compared to the testing of our initial population.

These testing results prove the ability and accuracy of our proposed artifact to be generalized and used in other FIBO ontologies and their concepts or even to be extended and applied using different data outlets. During the testing period, there exist some debatable subjects in NYC related to the context of new sources of revenue for the city, fixing the infrastructure, etc.

## 6. Evaluation of the final results and the artifact

Since our objective is to be able to scan any type of social media source and apply a scientific artifact which would enable us to better understand and extract some knowledge from messy unstructured data, we find that by using the publicly available FIBO ontology as the basis for our frames and slots, we are able to actually extract related meaningful knowledge. Our data results show that by organizing the unstructured data in this more structured way, we can extract more related tweets from Twitter with few false positives. The formal concepts and synonyms of FIBO provide structure and organization to the messy Twitter media, and this organization allows it to be joined to structured financial data that share the same concepts.

By applying different classifiers and tuning the search results, we discover that when using the best MTA classifier to classify the PANYNJ dataset, the results show the same number of tweets of 148 but a significant increase in the precision - 4.3% using the best MTA classifier compared to 17.8% false positives. Similarly, by using the best PANYNJ classifier to classify the MTA dataset, we get almost the same number of tweets of slightly above 400 with increase in precision of 4.7% false positives compared to 12%. Based on our current tests to date, it appears that both of the "best" classifiers perform similarly on both data sets. However, the MTA classifier demonstrates the more drastic improvement of false positive reduction when applied to the PANYNJ twitter feeds (13.5%) versus that of PANYNJ classifiers applied to MTA data (7.3%). It would be interesting to see if these results hold over time or if we might find a degradation or improvement of classifier performance.

Generally, there exist some limitations. For instance, in such environmental scanning we need more rich and voluminous datasets. We collected 10 months of data but believe that any subsequent additional tests (see the next section) might require a longer time frame. This will, of course, depend on the research topic. For instance, in the case of an event study using the type of artifact on social media presented in this paper, the time frame can be determined initially by using similar time frames used in event studies using quantitative analytic techniques. Like any big data project, the longer the period of data collection, the more instances we could extract, analyze, and subsequently achieve better accuracy (Abbasi et al., 2019). Furthermore, regarding

the PATH results above, many of the Tweets in the original feed created for PATH pick up the term "path" in the Tweet itself. This leads to phrases such as "revenue path" being classified and, consequently, often a false positive. We should consider a work-around for this issue in any future refinements of the artifact.

## 7. Conclusion and future research

This study is structured to capture tweets about past and pending economic events regarding PANYNJ and the MTA using an artifact developed from a slot and frame structure and the FIBO ontology. During the time period of this study, PANYNJ bond series were rated as stable Aa3 by Moody's and the Port Authority Board approved the application to raise funds to upgrade one of its airports (Airport-technology news.com, 2018). We anticipate that these economic events and other activities would provide sufficient Twitter data to demonstrate this type of artifact for environmental scanning.

Additionally, we anticipate that utilizing an artifact developed from a slot and frame structure and an ontology could facilitate the joining of "messy" unstructured data, such as that found in Twitter, to more structured financial data sharing the same FIBO concepts. The method of using the FIBO ontology in this manner should create a fuzzy match between data types that are ordinarily challenging and time-consuming to combine. The method can also be extended by including it along with more traditional regression and event study methodologies to allow for a richer data space in which decision makers can make judgements on bond analysis. It could also be combined with sentiment analysis of Twitter feeds. It should be noted that the semantics surrounding sentiment analysis, such as the attitude of riders about the MTA or PANYNJ, is not semantically equivalent to pulling information from the Twitter feed about bonds. Such studies must be careful to consider up front what the basis for any comparisons between the different methodologies will be. The validities of the comparisons must also be established and tested. The dimensions of such studies are such that they might need to be designed as research projects the results of which would be expected to be published in multiple research articles. Funding of such a project and the risk of failure to publish should be considered by any research group considering such an undertaking.

Although this study has been carefully grounded in DSR and accounting ontology theory, we should mention several assumptions made in our study that are typical for most examinations of social media. First, there is a mistaken assumption that Twitter feeds represent the true population. However, Twitter feeds only represent the tweets of the population that choose to actively interact on Twitter. Secondly, there are many subscribers of Twitter who do not actively post but may retweet and or just read tweets. Basically, most Twitter studies do not capture the tweets of the broad population, but only those of active Twitter participants. Thirdly, another assumption is that tweets represent the participant's actual meanings (semantic state). Twitter posts only display what participants elect to post, and as such could be abbreviated and/or modified. Some participants may feel comfortable posting in a manner similar to an unstructured "stream of consciousness" while others might post in a more measured and constrained manner. Some tweets may be more commercial or promotional in purpose. These issues point to the potential benefits of using a structured ontology such as that of FIBO for understanding a broad range of tweets that are of a more financial nature. However, since we are interested in Twitter streams for information extraction and not for sentiment understanding, it could be that these limitations are not as pervasive. That is, all it takes is one single tweet about an absent conductor on the Montclair Line to understand that there might be missing revenue – such an observation need only occur once, not multiple times.

Furthermore, Twitter might not be the medium with which users typically would choose to explicitly discuss municipal bond matters. They actually might tweet more frequently about concepts related to municipal bonds, such as "revenue". It is this conceptual context which allows the FIBO Conceptual Ontology method to shine. Using the FIBO Conceptual Ontology, it is possible to capture both the directly and indirectly related tweets. Many may argue that a simple keyword search of "municipal bond revenue" or other terms would suffice. However, such a search would not classify all the utterances and concepts in the Twitter data that are indirectly relevant, though captured with the FIBO ontology frames and slots. In fact, many users tweet about revenue in the context of their MTA/PANYNJ experiences and observations, but these tweets are not about bonds at all; yet, the tweets relate to the financial health of the bond-issuing agency. Here are a few such conceptually related tweets:

- @nyc311 Why are @MTA buses incapable of accepting dollar bills? Seems to me you are losing out on easy revenue!
- And of course there are not ticket collectors in sight. Riders suffer revenue suffers @NYGovCuomo @SenSchumer https:\/\/t.co\/S1Jpsd8CE3.

So even though the initial FIBO concept "revenue" does not relate exclusively to bonds (in the tweets captured it broadly relates to either the MTA, PANYNJ, NJPATH), it still relates indirectly to bonds in that revenues determine the overall financial condition of the bond-releasing agency and its bond performance and could potentially influence bond buying and selling behavior. It is not expected that users would tweet often about municipal bonds unless this topic is currently in the news. As mentioned earlier in the paper, the authors anticipate that the recent bond release would provide more instances for bond-related tweets than would normally be expected. Absent the recent bond release, there might have been zero explicit tweets about municipal bonds.

Additionally, there may be room for other FIBO ontologies to be applied to these Twitter data streams. Since many of the relevant tweets, particularly at the broad concept level of "revenue" could relate to municipal budgets or other financial contexts, it is hoped that interested parties could apply the artifact demonstrated here using other FIBO ontologies pertaining to other firms/municipalities. The method demonstrated here in this research should be generalizable across other FIBO ontologies and their concepts. This research demonstrates the potential for a FIBO ontology classifier artifact, when applied to messy social media feeds, to facilitate its fuzzy match to the relevant numbers in financial reports that share the same FIBO concepts. To illustrate, a bond analyst examining

the revenue streams of the MTA to arrive at a bond-buying decision may want to include a social media perspective of its revenues, such as that supplied by Twitter, as one of many sources of insightful information for an environmental scan. This analyst can then relate all "revenue" relevant tweets as identified by this FIBO ontology method to the revenue number in the financial report. In this case, the tweets and the number would share the same FIBO classifying concepts.

Secondly, this artifact could be applied to other sources of social media, such as Facebook postings and Instagram. It is anticipated that these data sources may be similarly challenging as Twitter to extract, classify, and match. Other more structured forms of social media such as blogs, articles, and reports may also provide rich resources for FIBO ontology and conceptually based classification and extraction. Furthermore, this artifact is not confined to the domain of municipal bonds. Other financial domains may be considered for which FIBO ontologies apply, in addition to any domain that is referenced with an ontology.

Furthermore, we did not consider the sentiments of the extracted tweets. In the future we may want to extend this research, adding additional nuanced information to our model by matching the sentiment of the relevant tweets to their relevant FIBO frames. Although previous research investigates sentiment and competitive analysis (He et al., 2013) and ontologies and sentiment analysis (Kontopoulos et al., 2013; Cotfas et al., 2015; Sánchez Rada et al., 2014), studies cannot be found that combine sentiment analysis and ontology-informed knowledge of Twitter feeds with their relevant financial or budgetary numbers. Loughran and McDonald (2019) discuss that as the applications of text mining broaden, researchers can capture more complex and relevant variables beyond traditional quantitative data.

Finally, if we are to consider this research as an application of representation or knowledge engineering, which is the intentional mapping of structured and unstructured data into a customized data architecture (Abbasi et al., 2019), there are risks that should be considered and avoided. First, the process should remain semi-automated where the extraction can be automated as a complimentary tool to irreplaceable human expertise. Domain, linguistic, and technical expertise are required to understand the application and relevance of certain tweets as representative knowledge of financial numbers. In this case, any automation would be best served to augment the human FIBO experts. Second, we may want to contextualize our findings where possible by seeking alternative sources of information regarding reported events. For example, if riders are jumping the gates at the subway stations, thereby not paying for fares which provide revenue, are there news reports online about subsequent remedies to this problem? It is quite possible that a more complete picture and context could be provided by accessing other sources of textual data. Third, as an application of knowledge engineering, it may be difficult to apply this specific ontology to other Twitter searches. Our current application of the FIBO ontology to Twitter data is representationally rich, incorporating highly problem-specific constructs and constraints tailored to our specific project. While the process may be generalizable, it is conceivable that our specific project ontology is not and might be best applied to other sources of textual data (Facebook feeds, Instagram posts, and so on) for the same specific purpose. In short, we may want to leverage other sources of complimentary text data to reduce the possibility of under or over fitting our model. Also, as we consider automating the maintenance of our domain-specific ontology (Hahn and Schnattinger, 1998), the same points apply. Our specific ontology will be incrementally updated as new concepts and terms are acquired from the tweets and other social media. Although this maintenance process will need to be formalized, it will still require domain, technical, and linguistic human experts.

The authors anticipate that this artifact would be of interest to bond issuers, regulators, analysts, investors, auditors, and academics. The study could be extended in multiple ways including those mentioned previously. Weighting individual Tweets by, for instance, the number of followers of the source could be added to the analysis. Adding Twitter threads instead of only individual Tweets might add some insights in particular business contexts. Other parts of the extensive FIBO ontology can be applied, including those of various expenses and other accounting classifications. It is possible that results of such studies might help to change FIBO by suggesting improvements of its underlying semantics. Other metrics of semantic strength or force of the Twitter stream such as the number of followers, likes and retweets could also be captured and tested. The method can also be extended to help develop the relationships between the semantics of ontologies and their expression in "local" language including the problems of synonyms and near synonyms, semantic shading and the development of meaning over time. Finally, exploring the development of Twitter feeds, including how long they should be and how to develop the initial key word filters may be an important topic as these methods gain currency.

This study contributes to accounting academic research since it develops an artifact for applying a formal business ontology (FIBO) to unstructured social feeds, such as Twitter, for knowledge representation. We envision a situation where either firm management or their auditors are looking to scan the firm's environment for potential risk or additional insights. Social media provides a rich potential source of environmental signals but also carries a lot of noise. The FIBO ontology was built by consulting with domain experts whose financial expertise has been captured in the ontology. Therefore, using such ontologies facilitates the identification and understanding of the nuanced threads in Twitter that pertain to such financial issues as probable and/or pending municipal bond releases. It is anticipated that these announcements would evoke a public reaction, and this study applies a formal accounting ontology to the processing of these social media extractions to facilitate and organize understanding. Using FIBO frames, one can capture more nuanced tweets that relate conceptually to the relevant financial data, a fuzzy match which might otherwise be overlooked in a less structured scanning approach.

## Acknowledgments

We wish to thank the University of Waterloo's Centre for Information Integrity and Information Systems Assurance, the participants and reviewers of our paper at the 11th Biennial Symposium on Information Integrity and Information Systems Assurance, and our paper's anonymous reviewers for all of their help and support. Zamil Alzamil thanks Deanship of Scientific Research at Majmaah University for supporting this work under Project No. 1441-142.

## References


Abbasi, A., Kitchens, B., Ahmad, F., 2019. The risks of AutoML and how to avoid them. available at: HBR.org. https://hbr.org/2019/10/the-risks-of-automl-and-how-to-avoid-them (October 24).

Airport-technology news.com, 2018. https://www.airport-technology.com/news/stewart-international-airport-new-york-undergo-renovation/ Available at:.

Aparaschivei, Florin, 2007. The importance of an accounting ontology. Econ. Inform. 1 (4), 5–10.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. J. Comput. Sci. 2 (1), 1–8.

Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P., 2010. Making words work: using financial text as a predictor of financial events. Decis. Support. Syst. 50 (1), 164–175.

Cotfas, L.A., Delcea, C., Roxin, I., Paun, R., 2015. Twitter ontology-driven sentiment analysis. New Trends in Intelligent Information and Database Systems. Springer, Cham, pp. 131–139.

Du, J., Zhou, L., 2012. Improving financial data quality using ontologies. Decis. Support. Syst. 54 (1), 76–86.

Feldman, R., Hirsh, H., 1996. Mining associations in text in the presence of background knowledge. KDD, pp. 343–346 (August).

Financial Industry Business Ontology (FIBO), d. https://www.edmcouncil.org/financialbusiness.

Gailly, Frederik, Laurier, Wim, Poels, Geert, d. Positioning REA as a business domain ontology. http://ideas.repec.org/p/rug/rugwps/07-460.html.

Gandhi, P., Loughran, T., McDonald, B., 2019. Using annual report sentiment as a proxy for financial distress in US banks. J. Behav. Finan. 1–13.

Geerts, G.L., 2011. A design science research methodology and its application to accounting information systems research. Int. J. Account. Inf. Syst. 12 (2), 142–151.

Geerts, Guido L., McCarthy, William E., 1999. An accounting object infrastructure for knowledgebased enterprise models. IEEE Intelligent Systems 1–6 (July/August).

Geerts, G.L., Graham, L.E., Mauldin, E.G., McCarthy, W.E., Richardson, V.J., 2013. Integrating information technology into accounting research and practice. Account. Horiz. 27 (4), 815–840.

Greco, F., Polli, A., 2020. Emotional Text Mining: customer profiling in brand management. Int. J. Inf. Manag. 51, 101934.

Guan, Jian, Cobb, Andrew, Levitan, Alan, 2006. An ontological model for accounting information systems. AMCIS 2006 Proceedings. http://aisel.aisnet.org/amcis2006/455 (Paper 455).

Hahn, U., Schnattinger, K., 1998. Towards text knowledge engineering. Hypothesis 1 (2).

He, W., Zha, S., Li, L., 2013. Social media competitive analysis and text mining: a case study in the pizza industry. Int. J. Inf. Manag. 33 (3), 464–472.

Hevner, A., March, S.T., Park, J., Ram, S., 2004. Design science in information systems research. MIS Q. 28 (1), 75–105.

Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N., 2013. Ontology-based sentiment analysis of twitter posts. Expert Syst. Appl. 40 (10), 4065–4074.

Kumar, B.S., Ravi, V., 2016. A survey of the applications of text mining in financial domain. Knowl.-Based Syst. 114, 128–147.

Lee, Thomas A., 2009. The ontology and epistemology of social reality in accounting according to Mattessich. Account. Public Interest 9, 65–72.

Liu, B., Zhang, L., 2012. A survey of opinion mining and sentiment analysis. Mining Text Data. Springer, Boston, MA, pp. 415–463.

Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: a survey. J. Account. Res. 54 (4), 1187–1230.

Loughran, T., McDonald, B., 2019. Textual Analysis in Finance (Available at SSRN 3470272).

Lupasc, Adrain, Lupasc, Ionna, Negoescu, Gheorghe, 2010. The role of ontologies for designing accounting information systems. The Annals of "Dunarea de Jos" University of Galati, Fascicle I. Economics and Applied Informatics (Years XVI, no. 1, ISSN 1584-0409).

March, S.T., Smith, G.F., 1995. Design and natural science research on information technology. Decis. Support. Syst. 15 (4), 251–266.

Mattessich, Richard, 2003. Accounting representation and the onion model of reality: a comparison of Baudrillard's orders of simulacra and his hyperreality. Acc. Organ. Soc. 28 (5), 443–470.

McCarthy, W.E., 1982. The REA accounting model: a generalized framework for accounting systems in a shared data environment. Accounting Review 554–578.

McCarthy, W.E., 2012. Accounting craftspeople versus accounting seers: exploring the relevance and innovation gaps in academic accounting research. Account. Horiz. 26 (4), 833–843.

Mellouli, S., Bouslama, F., Akande, A., 2010. An ontology for representing financial headline news. Web Semant. Sci. Serv. Agents World Wide Web 8 (2–3), 203–208.

Moffitt, K., Vasarhelyi, M.A., 2013. AIS in an age of big data. J. Inf. Syst. 27 (2), 1–19.

Mostafa, M.M., 2013. More than words: social networks' text mining for consumer brand sentiments. Expert Syst. Appl. 40 (10), 4241–4251.

Murthy, Uday, Geets, Guido L., 2017. An REA ontology-based model for mapping big data to accounting information systems elements. J. Inf. Syst. 31 (3), 45–61.

Pak, A., Paroubek, P., 2010. Twitter as a corpus for sentiment analysis and opinion mining (In LREc). 10 (2010), 1320–1326.

Sánchez Rada, J.F., Torres, M., Iglesias Fernandez, C.A., Maestre Martínez, R., Peinado, E., 2014. A Linked Data Approach to Sentiment and Emotion Analysis of Twitter in the Financial Domain.

Saura, J.R., Bennett, D.R., 2019. A three-stage method for data text mining: using UGC in business intelligence analysis. Symmetry 11 (4), 519.

Searle, J.R., 1995. The Construction of Social Reality. Free Press, New York, NY.

Stutzman, Fred, 2007. The 12-minute definitive guide to Twitter. https://web.archive.org/web/20080704074026/http://dev.aol.com/article/2007/04/definitive-guide-to-twitter. (Accessed 10 February 2018). (April 11).

Syed, A., Gillela, K., Venugopal, C., 2013. The future revolution on Big Data. Int. J. Adv. Res. Comput. Commun. Eng. 2 (6), 2446–2451.

Wang, S., Zhe, Z., Kang, Y., Wang, H., Chen, X., 2008. An ontology for causal relationships between news and financial instruments. Expert Syst. Appl. 35 (3), 569–580.

Wang, S., Xu, K., Liu, L., Fang, B., Liao, S., Wang, H., 2011. An ontology based framework for mining dependence relationships between news and financial instruments. Expert Syst. Appl. 38 (10), 12044–12050.

Warren Jr., J.D., Moffitt, K.C., Byrnes, P., 2015. How Big Data will change accounting. Account. Horiz. 29 (2), 397–407.