View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by AIS Electronic Library (AISeL)

Sentiment Analysis Meets Semantic Analysis: Constructing Insight Knowledge Bases

Completed Research Paper

Zirun Qi
Georgia State University
Atlanta, GA, USA
zqi1@gsu.edu

Veda C. Storey
Georgia State University
Atlanta, GA, USA
vstorey@gsu.edu

Wael Jabr
Georgia State University
Atlanta, GA, USA
wjabr@gsu.edu

Abstract

Numerous Web 2.0 applications collect user opinions, and other user-generated content in the form of product reviews, discussion boards, and blogs, which are often captured as unstructured data. Text mining techniques are important for analyzing users' opinions (sentiment analysis) and identifying topics of interest (semantic analysis). However, little work has been carried out that combines semantics with user's sentiments. This research proposes a Sentiment-Semantic Framework that incorporates results from both semantic and sentiment analysis to construct a knowledge base of insights gained from integrating the information extracted from each type of analysis. To evaluate the framework, a prototype is developed and applied to two different domains (e-commerce and politics) and the resulting insight knowledge bases constructed.

Keywords: Sentiment analysis, semantic analysis, knowledge base, ontology, affect theory, text mining, user-generated content, insight knowledge base

Introduction

"What other people are thinking" is always an essential piece of information during the decision-making process in business intelligence (Pang and Lee 2008). To gain such insights, both sentiment analysis and semantic analysis are advancing and being adopted in marketing, e-commerce, online communities, social media, and other applications, due to the continued, explosive growth of user-generated content on the Internet. The accessibility of large and variant user-friendly Web 2.0 features enable users to share their experiences, which are often represented online as unstructured data. However, this data, properly mined, has the potential to provide insights into customer purchase habits, reactions, interest levels, and other such behaviors. For example, by performing sentiment analysis of consumer purchase behavior and product feedback, one can attempt to predict both positive and negative reactions from other customers on certain product features, without the need for complete information about such features (Liu et al. 2005).

A richer approach to extracting and representing valuable information, involves appreciating the inherent semantics of user-contributed online content, in addition to sentiment mining. Suppose, for example, a customer wants to know whether a specific model of a smartphone can be read easily in sunlight. Customer reviews might indicate that the screen of this smartphone received 45% positive and 15% negative feedback comments. However, this does not provide information on the customer desires, namely, screen performance in sunlight. The customer must still manually search the reviews on keywords related to screens. Adding such manually-extracted information into a knowledge base would be valuable for others (e.g., marketing managers or potential customers).

With respect to semantic analysis, the Semantic Web is intended to organize web resources in a form that can be universally shared, appearing as "subject-predicate-object" expressions (Berners-Lee et al. 2001) and intended to be a significant improvement in the use of the web (Fazzinga et al. 2011). Semantics are captured by mapping to ontologies, of which there are varying amounts of expressiveness (Hendler and Golbeck 2008). In sentiment analysis, Multi-Perspective Question Answering (MPQA) is employed. The Subjectivity Lexicon categorizes sentiment information as "positive or negative," which are polarity labels (Wilson et al. 2005). Several research efforts attempt to combine semantic and sentiment analysis results (e.g. Liu et al. 2005; Qiu et al. 2011). However, these projects focus primarily on the accuracy of sentiment analysis results (in polarity), rather than integrating the results and representing them in an insightful and reusable knowledge base.

The question addressed in this research then becomes: *Is it possible to combine sentiment analysis and semantic analysis to mine unstructured data and use the results to build an insightful, reusable knowledge base of results?* This research takes a design-science approach by: (1) adapting affect theory as kernel theory (Walls et al. 1992) to design a framework, (2) developing a prototype and testing it with real world user-generated content, and (3) evaluating both the framework and the prototype (Baskerville et al. 2015). The specific objective of the research is to: develop a Sentiment-Semantic Framework and a prototype for capturing and representing both semantics and sentiment of user-provided online content. The contribution of doing so is to provide an approach to integrating sentiment and semantic analysis, based upon ontologies and affect theory, to support both individual and organizational level decision making.

Related Research

Research in computer science, linguistics, psychology, and information systems all approach sentiment and semantics quite differently. In linguistics, semantics is regarded as the study of meaning of human expression through language (Ullmann 1962), whereas computer science studies regard semantics as a knowledge representation issue (Guarino and Giaretta 1995). The Semantic Web has an objective of enabling machines to "*comprehend* semantic documents and data, not human speech and writings" (Berners-Lee et al. 2001). This section first provides an example and, then, an overview of related concepts for this research, which provide the foundation for the proposed framework.

Consider Amazon.com's relatively new review summary feature that attempts to provide an aggregated, high-level assessment of reviews as shown in Figure 1. Three statements are listed to the right of the "stars". Each statement describes one or more features of the reviewed product. However, the exact

features are not extracted or aggregated. The first statement describes "weight," "screen light," and "font;" the second describes "weight," "screen light," and "touch screen" features. These features, though, are not explicitly stated. Nor do they represent the common and aggregated features (such as "weight" and "screen"). Sentiment information is aggregated at the sentence level; however, feature level sentiment information (e.g. very good lighted screen) is not extracted.

Figure 1. Amazon.com customer reviews

Semantic and Sentiment Analysis

To represent human affective information, we adopt the term "semantics" to represent real world objects, facts, actions, events, and people (or more conceptually, real world entities); and "sentics" as identical to "sentiment," to represent human affective information (Cambria and White 2014). Semantics (real world entities) are objective, and thus, independent of human feelings. In contrast, sentiments capture subjective, human feelings about real world entities. For example, the sentence "the CPU in this computer can run at 3Ghz clock rate," describes real word objects and fact. Nevertheless, different people can have different feelings about the same computer's performance, usability, etc., based upon their individual experience.

Polarity results for sentiment analysis are usually in the form of a proportion for positive/negative, providing only one-dimensional information in sentiments. Compared to face-to-face communication, this analysis provides the least rich approach to capturing human affective information. To improve the quality of sentiment analysis results, based upon the level of information richness (Daft and Lengel 1983), we examine literature in psychology for related approaches. Affect theory is a practical way to expand single dimensional polarity results to a richer level (Scherer 2005). Affect refers to the feelings of an individual based upon his or her experience and can represent a type of emotion, graduation, orientation, or polarity (Frijda 2007; Garcia-Crespo et al. 2010; Barsade and Gibson 2007). Thus, it can highly influence the content provided by a user. Affect theory organizes human affective information into discrete affect categories, and links each category to its common emotional responses. By incorporating affect, the sentiment information in user-generated data can be coded into different categories, resulting in a multi-dimensional approach.

Ontology and Knowledge Base

Also relevant to this research is the concept of an ontology defined as "as explicit specification of a conceptualization" (Gruber 1993) for the purpose of knowledge-based systems development. The ontology field is defined as a "Formal Ontology," which is "not so much the bare existence of certain objects" (as defined in Philosophy), but rather, "the rigorous description of their forms of being, i.e. their structural features" (Guarino and Giaretta 1995). We adopt the following definition of "ontology" (with lower case letter "o") in this research: "a logical theory which gives an explicit, partial account of a conceptualization" (Guarino and Giaretta 1995). In practice, an ontology distinguishes and merges real world entities as hierarchical relationships based upon their characteristics.

An ontology is closely related to knowledge representation and knowledge acquisition (Guarino and Giaretta 1995). A knowledge base in the knowledge engineering field is "a result of modelling activity

whose object is the observed behavior of an intelligent agent embedded in an external environment" (Clancey 1993). In addition, a knowledge base refers to an objective reality instead of being in an agent's "mind". However, for any situation in which human "minds" are involved, sentiment is part of the essence of decision making. This research, then, combines sentiment information and semantics to create what we call an "insight" knowledge base, which is intended to capture valuable knowledge for reuse that captures these combined results.

In Natural language processing (NLP) evolution, both sentiment analysis and semantic analysis are considered part of NLP, with sentiment (sentic) as human affective information considered to be the "key for common-sense reasoning and decision making" (Cambria and White 2014). Suppose, for example, we have the affect information: "the CPU in this computer can run at a 3 GHz clock rate, but I still feel it is very slow". "Computer" and "CPU" are two real world entities, and appear in a 2-level hierarchical relationship: Computer -> (has) CPU. The human affective information, "feel slow," could be different based upon one's past experience. Therefore, for more insightful analysis, we need to capture and aggregate different affect information of a human being, and combine it with the semantic analysis result. Table 1 defines the terms used in this research.

Table 1. Definitions

Term	Definition
Semantic	Representation of real world entities, more specifically: real world objects, facts, actions, events, and people (Cambria and White 2014)
Sentiment	Representation of human affective information (Cambria and White 2014).
Affect theory	Organization of human affective information into different categories (Frijda 2007)
Ontology	Explicit specification of a conceptualization; distinguishes and merges real world entities as hierarchical relationships based on their characteristics (Gruber 1993).
Knowledge Base	A result of a modelling activity whose object is the observed behavior of an intelligent agent embedded in an external environment; objective reality (Clancey 1993).
Insight Knowledge Base	Knowledge Base augmented with human affective information.

Table 2 summaries relevant, prior research on semantic and sentiment analysis. As can be seen from the table, all of the sentiment analysis results focus on polarity (positive/negative), which is easy to aggregate. However, this approach lacks the richness of human affective information. Moreover, with respect to semantics, most research is limited to semantic analysis applications. Abbasi and Chen (2008), for example, perform topic classifications on emails and analyze the semantics of the topics. Liu et al. (2005) attempt to combine semantic analysis with sentiment, but they too take a polarity approach. Although there are limited semantic analysis applications, information systems has made more attempts to perform sentiment analysis by techniques such as Part-Of-Speech (POS) tagging and lexicon tagging (Toutanova et al. 2003). However, polarity results lack information richness, which is the essence of decision making (Cambria and White 2014).

Research Methodology

This research develops a Sentiment-Semantic Framework for the analysis of online user-generated content that combines semantic and sentiment analysis to build an "insight" knowledge base of extracted information. Figure 2 depicts the framework, which includes the relationships among sentiment analysis, semantic analysis, ontology, and knowledge base. The components are described below.

Raw and Unstructured Text: User-generated content is extracted from the internet. For the purposes of this research, the data is represented in text format. However, the framework should be extendable to other formats, such as audio and video.

Table 2. Semantic and Sentiment Analysis from Prior Research

Study	Data	Domain / Goal	Semantic Analysis Technology	Result	Sentiment Analysis Technology	Result
Turney and Littman (2003)	General web pages	General / Analysis sentiment in different context	None	None	Lexicon tagging, POS tagging, Semantic oriented approach	Polarity
Das and Chen (2007)	Online forums	Finance / Prediction of stock index movement	None	None	Lexicon tagging, Classification	Polarity
Archak et al. (2011)	Online consumer reviews	e-Commerce / Prediction of sales	Crowdsourcing for product feature	List of product features	POS tagging, Lexicon tagging, Clustering, Crowdsourcing	Polarity
Abbasi and Chen (2008)	Company internal emails	Text analysis / Classification of topic, opinion, style, genre	Topic Categorization	Topic Clusters	Writeprints, Ink Blots	Polarity
Chau and Xu (2012)	Online blogs	e-Commerce / Business Intelligence	None	None	Manually Classification	Polarity
Doan et al. (2002)	Commonsense Knowledge	General / Create accurate semantic mappings	Distribution-based Similarity Measures	Ontologies	None	None
Liu et al. (2005)	Product Reviews	e-Commerce / Business Intelligence	POS tagging, Short sentence segments	List of product features in different levels	POS tagging, Lexicon tagging	Polarity combined with semantic result

Semantic Axis: This axis describes the level of the ontology. A low ontology refers to a very specific domain ontology with limited entity coverage. A high ontology refers to a wider and deeper ontology and may cross domains (Noy and McGuinness 2001). For instance, "iPhone screen -> Resolution" is a low detailed ontology representation. "Smartphone -> iPhone -> Screen -> Resolution" is a higher ontology representation.

Sentiment Axis: This axis describes the level of the dimensional information. A positive/negative polarity result is a lower representation (one dimension) compared to a multiple dimensional result such as "anger/happy/sad/satisfied" affect categories.

Approach 1 and 2: Depending on the analyzing sequences, there are two ways to perform semantic and sentiment analysis. Approach 1 extracts the entity information and organizes them into a hierarchical structure first, and then incorporates the sentiment information for each of the entities. Approach 2 implements the opposite order.

Insight Knowledge Base: An insight knowledge base is generated that associates the aggregated human affective information (sentiments) with corresponding real world entities (semantics). It provides

more usability and generalizability across different organization environments than a traditional knowledge bases (e.g., those for the Semantic Web), that represent only the relationships among real world entities. Because of the additional sentiment information, the insight knowledge base captures knowledge that can be used (and reused) to make better decisions. This could be especially useful in business intelligence applications.

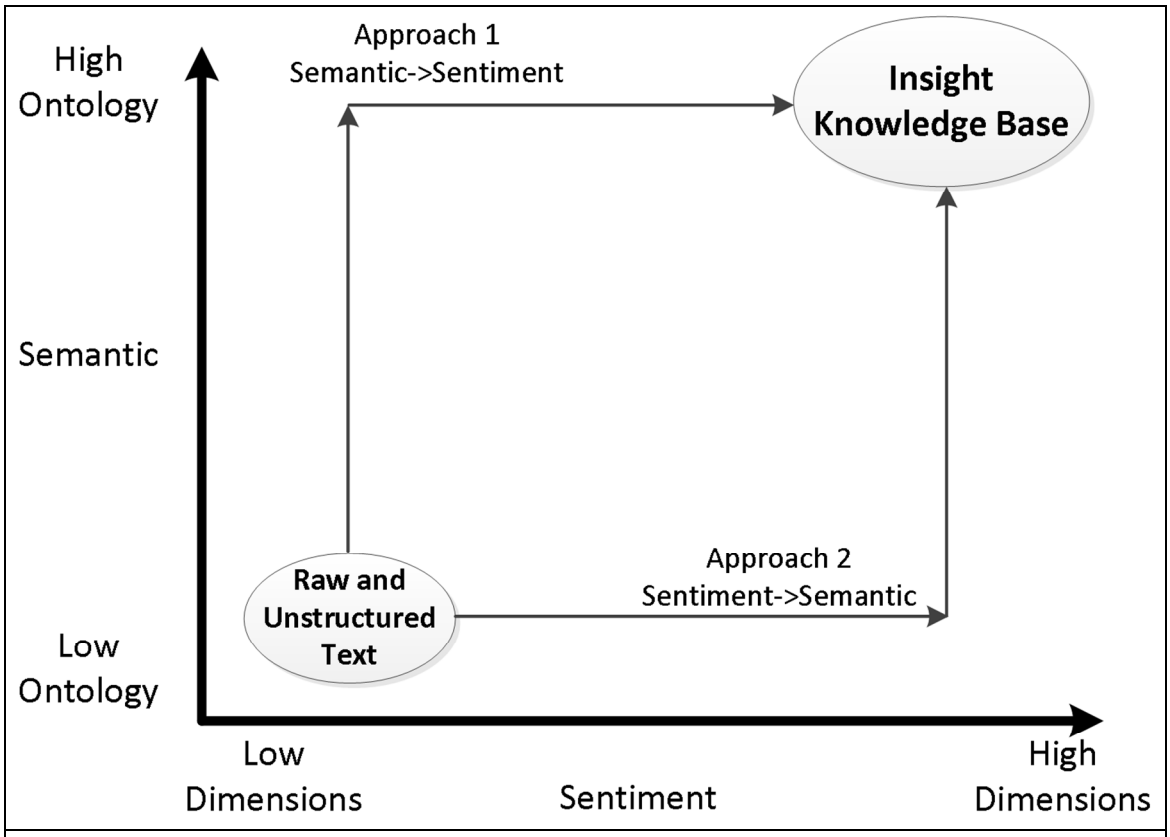

Figure 2. Sentiment-Semantic Framework

Conceptual Model of Database Design: The post-process is to store the insight knowledge base into a database for reusability, scalability, and comparability. An insight knowledge base, as structured data, can be easily transformed into a relational database. Figure 3 shows the entity-relationship model as an insight knowledge base design.

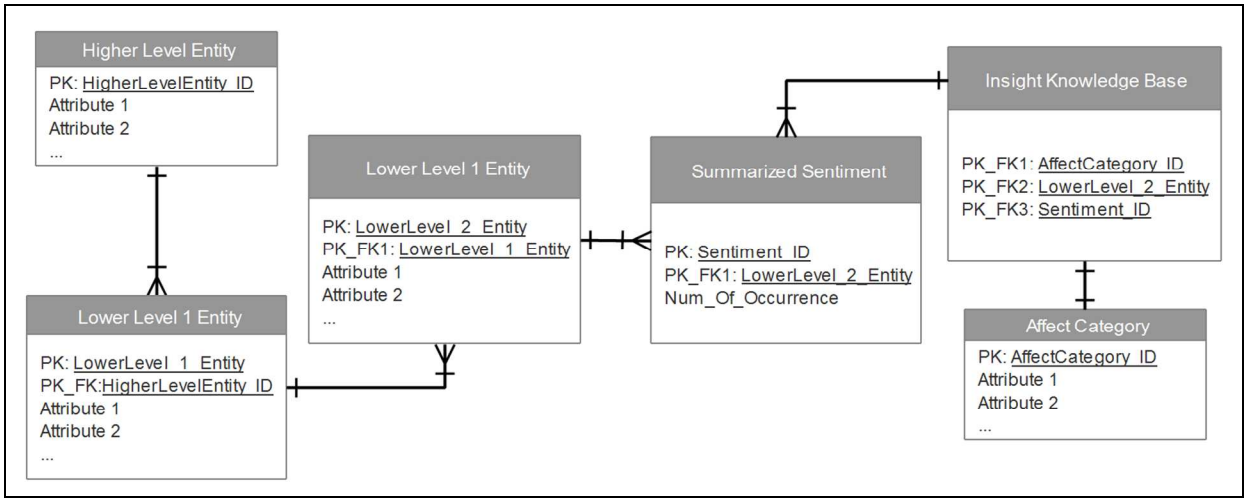

Figure 3. Entity-Relationship Model of Insight Knowledge Base Design

Artifact Description and Implementation Result

This research adopts a design science research approach with a framework artifact (March and Smith 1995; Hevner et al. 2004). To test the framework, it is implemented in a prototype, the architecture of which is shown in Figure 4.

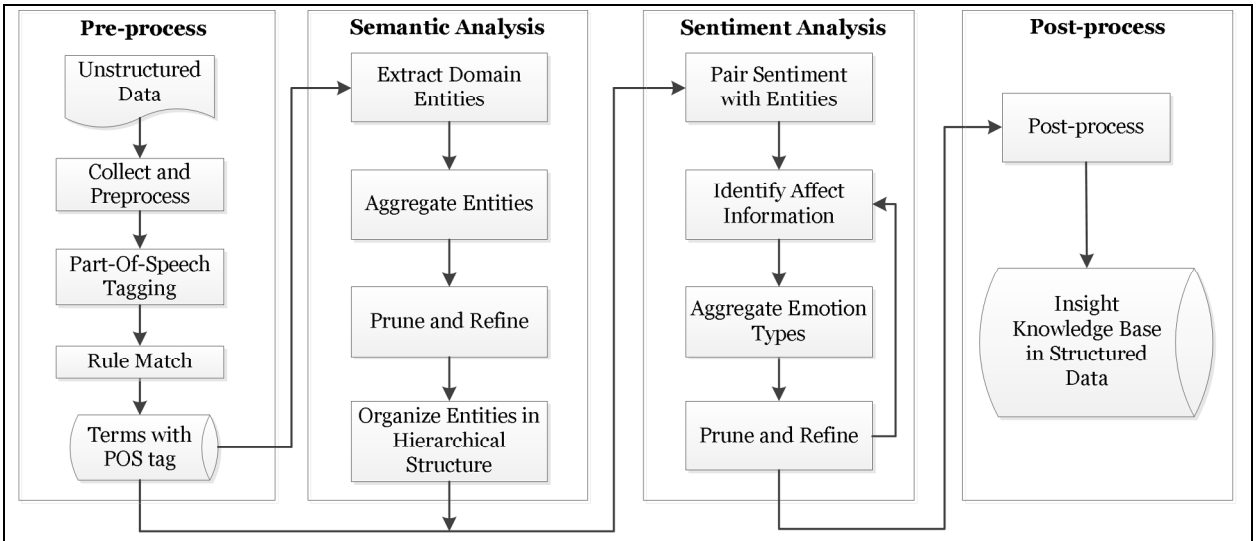

Figure 4. Prototype Architecture

The prototype processes and transforms unstructured data into a structured knowledge base of insights gained. All of the processes have been implemented using the KNIME Analytics Platform (KNIME.org). Although the overall objective is to completely automate this process, some parts of the steps still require manual effort.

Data Source: For illustration and evaluation purposes, two data resource are selected from two different domains as input to the prototype. The first consists of 500 customer reviews on Amazon.com for the product "Kindle Fire HD 7". The second dataset consists of 500 threads (3640 posts) from a US political discussion forum (usmessageboard.com) on the topic of illegal immigration.

Data Collection and Pre-processing: The webpage crawler collects user-generated content (unstructured or semi structured data), particularly in the form of user reviews (Amazon.com) and forum posts (usmessageboard.com). The pre-process algorithms clean the content to remove extra punctuations, blanks, URLs, numbers etc. Then, each piece of content is split into sentences. The Stanford NLP POS tagging tool (Toutanova et al. 2003) identifies nouns, verbs, adjectives, adverbs and other parts of speech. The tagged sentences are then passed to the rule matching process to identify short terms of interest. Table 3 summarizes the rules for doing so. The resulting short terms are then stored into a database for further analysis.

Semantic Analysis: From the POS tagger rule matching result, the terms with the POS tag are first used to identify and extract domain entities. For example, "access" and "service" will be extracted from "Easy access to amazon service" in Table 3. The entities with the same meaning but found in different forms, such as singular/plural and synonyms, are aggregated. Furthermore, some implicitly expressed semantics are extracted and linked to entities, such as "is too big" refers to "size". Then, entities that are not within the scope of the domain are pruned and refined. For example, in the semantic analysis result of Kindle fire, iPad appears as an entity, but is actually not part of a sub domain of Kindle fire. Finally, the domain entities are organized into a hierarchical structure, with part of this process performed manually.

Table 3. Examples of Rules Applied in POS Tags Rule Matching	
Rule	**Example**
Adjective -> 0 to 3 words -> Noun	Easy access to amazon service
Noun-> 0 to 3 words -> Verb -> 0 to 3 words -> Adjective	Battery life is extremely short
Noun -> 0 to 3 words -> Noun	Screen are smaller than iPad
Verb -> 0 to 3 words -> Noun/Adjective	Have unlimited cloud storage
"->" means followed by	

Sentiment Analysis: After the hierarchical structure of the entities is created, the results of the rule matching are double-checked by filtering out those short terms that only include created entities. The purpose of doing so is to identify and extract sentiment words paired with entities. For example, in "Battery life is extremely short", the word "short" is perceived by the user as sentiment information, and paired with the entity "battery." Furthermore, the sentiment words are sorted by the same entity, and summarized by similar meanings (e.g. the words "short" and "low" have similar meaning to describe battery capacity). Then, based on the word stems and affect category list (Scherer 2005), we develop an affect lexicon to map sentiment words into a corresponding affect category. Word stem completion is carried out to obtain the full words in the affect category list. Then, the synonyms of each full word are extracted from Merriam-Webster dictionary. Finally, the semantic-sentiment "pairs" are pruned and refined to ensure that the categorization is appropriate and aggregates the sentiment information by count within the same affect category. This allows us to combine and aggregate sentiment and semantics at the same time. In addition, it provides the capability to make comparisons across different domain ontologies (e.g. different products or different discussion topics).

Post-processing and Output: Aggregated semantic and sentiment information is represented as an entity-relationship model (Chen 1976), so the results can be stored and linked together for further retrieval and analysis.

Insight Knowledge Base: The results are represented in a structured knowledge base, called the insight knowledge base, the purpose of which is to store results that can be reused. For example, the prototype organizes entities into a detailed hierarchical structure. It also captures the paired sentiment analysis results at two levels: summarized sentiment; and affect category.

Automation: Previous studies demonstrate that not all the processes in our prototype can be performed in full automation. Liu et al. (2005), for example, provide only semi-automated tagging of product reviews for both product features (i.e. semantics) and sentiments. Table 4 summarizes the extent of automation in our prototype.

In the semantic analysis, Prune and Refine, involves a manual effort to merge similar entities for the same concept and add entities from implicit semantics. For example, "This smart phone is too big" is an implicit expression of its size. The process Organize Entities needs manual effort to correctly arrange entities in a hierarchical structure.

In the sentiment analysis, Identify Affect Information, requires some manual effort when non-adjective words contain sentiment information. For instance, "This smart phone does not fit my pocket" indicates a piece of sentiment information. The Prune and Refine process focuses on the words with sentiment information categorized into the correct affect category, based upon the affect category lexicon. For example, "the screen is shining" can refer to difficulty seeing in the presence of bright light. However, instead, it can be intended as a compliment to the screen.

Table 4. Automation in Prototype		
Process	**Automation**	**Explanation**
Pre-processing	Full	Web crawler, POS tagging tool, and rule matching are fully automated by text mining tools.
Semantic Analysis		
Extract Domain Entities	Full	Extraction of all nouns based on POS tagging result.
Aggregate Entities	Full	Aggregation by the same entity.
Prune and Refine	Partial	Manual effort is needed to merge similar entities, and add more entities from implicit semantics.
Organize Entities	Partial	Manual effort helps organizing entities to construct a correct hierarchical structure
Sentiment Analysis		
Pair Sentiment with Entities	Full	Pairing process is based on the result of POS tagging within the same term.
Identify Affect Information	Partial	Although the POS tagger of adjective words can be recognized as sentiment information, words in some other POS tagger also carry the sentiments. Also implicit expression of sentiments needs manual effort.
Aggregate Affect Types	Full	Affect types are aggregated by affect categories.
Prune and Refine	Partial	Manual effort needs to check if affect types are correctly categorized.
Post-processing	Full	Insight knowledge base is stored in a predefined relational database.

Results

Table 5 shows the top 20 entities as semantic analysis resulting from applying the prototype to a set of "Kindle Fire HD 7" reviews from Amazon.com. Based on 500 product reviews, the related entities, such as functions, components, capabilities, are extracted. The top 20 frequently occurring entities within this product domain are shown. Entities that appear in different forms (e.g., singular/plural as in book/books), or represent the same concept (e.g. problem/issue), are merged. The full entity list represents the lower level ontology of this product (i.e. Kindle Fire HD 7).

Table 6 shows the insight knowledge base of "Screen," as one entity in the ontology associated with this product. The left two columns are expanded semantic analysis results. The detailed entities in the next lower level ontology are extracted from terms using POS tag. The paired sentiments are summarized based upon the detailed entities, and also counted by the number of occurrence. Then, each summarized sentiment is mapped to one affect category, based upon the affect category list from Scherer (2005). In addition, all affect categories are labelled with their polarity result (positive/negative) based on MPQA Subjectivity Lexicon (Wilson et al. 2005), to be consistent with previous sentiment analysis research (e.g. Abbasi and Chen 2008; Chau and Xu 2012; Liu et al. 2005).

Table 5. Semantic Analysis Results (Entity Extraction) of Product Reviews from Amazon.com			
Entity	**Num. of Occurrence**	**Entity**	**Num. of Occurrence**
Books	76	Purchase	28
Apps	64	Size	26
Time	63	Battery	25
Screen	53	Power	24
Access	42	Problem	24
Music	36	Display	21
Internet	32	Button	20
Movies	32	Quality	20
Price	30	Wi-Fi	20
Video	29	Software	18

Table 6. Example of Insight Knowledge Base of Product Reviews from Amazon.com (Semantic Analysis Combined with Sentiment)					
Semantic		**Sentiment**			
Entity (Ontology)	**Detailed Entity (Ontology)**	**Summarized Sentiment**	**Num. of Occurrence**	**Affect Category**	**Polarity Result (Pos/Neg)**
Screen (of Kindle Fire HD 7)	Touch screen	Sluggish	13	Disappointment	Negative
		Useful	11	Contentment	Positive
	Screen size	Small	7	Disappointment	Negative
		Large	3	Contentment	Positive
	Screen quality	Vivid	10	Pleasure	Positive
		Glare (reflex light)	9	Irritation	Negative
		Total	53		24/29 (Pos/Neg)

The total number of occurrence of summarized sentiment is 53 (Table 6), which equals the number of occurrence of the entity "Screen" (Table 5). This indicates that, for every occurrence of "screen" mentioned in the reviews in our analysis, there is a piece of sentiment information associated with it.

Finally, all of the other entities shown in Table 5 are analyzed by the same way as "Screen" in Table 6. Based upon the insight knowledge base in Figure 3, an instantiation of a conceptual model (i.e. an entity-relationship diagram) is also implemented.

Evaluation

Following the classification of artifacts proposed by Venable et al. (2014), the artifacts were evaluated in two steps. The first artifact is the Sentiment-Semantic Framework shown in Figure 2. This framework is a process artifact, which could be used as a guideline for other researchers and practitioners to construct an

insight knowledge base or modify an existing semantic (only) knowledge base. The second artifact is the prototype, which is a product artifact involving human interaction when a decision is made. For this, we evaluate the prototype and its resulting utility (Venable et al. 2014).

Evaluation of Framework

Based upon the design science research (DSR) evaluation strategy selection framework of Venable et al. (2014), the Sentiment-Semantic Framework can be placed in the "Artificial-Ex Ante" quadrant for the following reasons. (1) It is a purely technical artifact without any human interaction. (2) It has little conflict among different evaluation criteria. (3) It has lower cost and is faster, since there is no instantiation created in the framework design process. Then, the criteria-based evaluation is chosen from the "Artificial-Ex Ante quadrant." To address our research question, four components are needed to design the framework as shown in Table 7. The evaluation results show how the Sentiment-Semantic Framework answers the research question.

Table 7. Criteria-based Evaluation of Sentiment-Semantic Framework

Criteria to Answer Research Question	**Framework Evaluation**
How to deal with user-generated unstructured data?	Raw data is decomposed into semantic and sentiment parts based on the definition (i.e. real world entities vs. human affective information).
How to improve semantic analysis and ontology?	Based on definition of ontology, a more detailed ontology is the way to obtain a better semantic result.
How to improve sentiment analysis to get more information richness?	Based on affect and information richness theory, a higher dimension sentiment analysis result has a more precise representation of human affect.
How to combine semantic and sentiment results? And why an insight knowledge base is better?	Both Approach 1 and 2 indicate that sentiment and semantic information are kept together in the analysis process. Final result is an improved knowledge base because combined sentiment and semantic information is a key to decision making.

Evaluation of Prototype

Comparing to the framework, the design process of the prototype is more focused on real problems and users, considering the capabilities of current technology. With the example of the insight knowledge base from real world data, the prototype is identified as "Naturalistic- Ex post" quadrant in the DSR evaluation strategy selection framework (Venable et al. 2014). In addition, it has: (1) an instantiation software and real users; (2) the highest cost (time to develop) and highest risk (results may not be good enough), and (3) the capability to show side effects (weaknesses). Then, a case study evaluation is selected (Venable et al. 2014).

Although the previous section shows the insight knowledge base created for one real world example (Kindle Fire HD 7), it provides only proof-of-concept. Then a different dataset was collected from a totally different domain (US political discussion forum), for evaluation purposes.

Similar to Table 5 and 6, Table 8 shows the top 20 entities as the results of the semantic analysis (entities extraction) from the political discussion board. Table 9 shows the lower level and detailed entities within the "Illegals" entity, which is associated with summarized sentiments, number of occurrences, and the affect category as the insight knowledge base.

Table 8. Semantic Analysis Result (Entities Extraction) of Discussion from US Political Forum			
Entity	**Num. of Occurrence**	**Entity**	**Num. of Occurrence**
American(s)	342	Citizenship	138
Law(s)	326	Illegals	137
Country	260	Years	133
People	259	Amendment	129
Immigration	238	America	128
States	186	Border	123
Immigrants	183	Problem	118
Time	152	Children	116
Aliens	150	Government	111
Citizens	141	Work	110

Table 9. Example of Insight Knowledge Base of Discussion from US Political Forum (Semantic Analysis Combined with Sentiment)					
Semantic		**Sentiment**			
Entity (Ontology)	**Detailed Entity (Ontology)**	**Summarized Sentiment**	**Num. of Occurrence**	**Affect Category**	**Polarity Result (Pos/Neg)**
Illegals	Work/Jobs	Unlawful	9	Guilt	Negative
		Harmful	5	Fear	Negative
		Cheap	4	Contempt	Negative
		Useful	4	Contentment	Positive
		Impossible	2	Desperation	Negative
	Deport	Agreeable	8	Positive	Positive
		Against	3	Negative	Negative
	Giving Citizenship	Useful	3	Contentment	Positive
		Funny	1	Amusement	Positive
		Progressive	1	Enthusiasm	Positive
	Criminals	Risky	5	Fear	Negative
	Cost of Medical	Staggering	2	Tension	Negative
	Education	Clogging	1	Anxiety	Negative
		Lacking	1	Anxiety	Negative
		Total	49		17/32 (Pos/Neg)

For semantic analysis, considering the results from both the Amazon.com and the US politics cases, we can conclude that the prototype can identify real world entities and their corresponding ontologies from

different real world domains. It can also extract lower level entities to construct a hierarchical structure as an ontology. Our semantic analysis results are comparable to previous studies (Doan et al. 2002, Liu et al. 2005). Additional improvement has been shown in the sentiment analysis results. Instead of aggregating positive/negative results, our prototype shows aggregated summarized sentiment words with the number of occurrences. Furthermore, the corresponding affect category is labelled to each summarized sentiment term.

To address our research question, the prototype closely connects semantic and sentiment analysis resulting from the beginning (pre-process) through the end (insight knowledge base). With the terms of POS tags retained as a middle result, the separation of summarized sentiment words and affect category provide the "trace-back" capability at different levels. Despite the variation of sentiment information, the insight knowledge bases have a unified affect category list, so they can be compared and reused across domains.

Discussion

The Sentiment-Semantic Framework proposes an integrated sentiment and semantic analysis that relies on ontologies, knowledge bases, and affect theory to build a useful and insightful knowledge base, intended to support both individual and organizational level decision making. The implementation and application of the framework for incorporating semantic and sentiment analysis demonstrate reasonable results. The intention of the resulting insight knowledge base is to provide a unique way to capture and represent useful knowledge extracted from the integration of semantics and sentiment, so it can be reused and applied to multiple applications across different domains.

Advantages

This research separates the artifact design into a framework and prototype, which is intended to have several strengths. (1) Clarity: before diving into the design process, definitions of semantic, sentiment, knowledge base and ontology, based upon the literature, were carefully reviewed and then adopted. (2) Generalizability: the framework provides guidelines for how to solve general or specific domain text mining and insight knowledge base construction problems. (3) Technology isolation: the separation isolates current technology and leads towards a future direction for natural language processing development and more insightful knowledge base construction.

Challenges and Limitations

A number of challenges remain. As identified by Metzler and Croft (2007), latent concepts should also be captured. For example, the product features (entities) in the sentence such as "I was hoping that the fire had a screen like the less expensive kindles" cannot be extracted by our prototype. We know that "the fire" means kindle fire, and "screen" refers to the screen of a kindle fire. The concept of "less expensive kindles" cannot be clearly identified based on an entire review, because "less expensive kindles" refers to several other kindle products, whose price or screen (ontology) may not be shown within this single kindle domain. Addressing this type of problem, may require further extraction mechanisms or more useful and complete ontologies. The insight knowledge bases produced may also require restructuring before they can be applied.

With respect to sentiment analysis, a similar challenge emerges when the same sentiment words appear within different contexts/domains. For example, "small screen" may show totally opposite sentiments in laptop and smartphone reviews. A small screen for a laptop is likely to represent more portability, but more likely to represent insufficient screen size for a smartphone. One possible solution is to consider all co-occurrences of other sentiment words with "small screen" to create a list of likely, correct sentiment representations for further analysis. As proposed by Cambria and White (2014), human common-sense knowledge is the key to correctly decomposing natural language text into sentiment information in different contexts. Co-occurrence of a nested extraction of "sentiment word" with "sentiment-semantic terms" (e.g. a more portable small screen) will enrich an insight knowledge base. However, the computational cost will increase dramatically. As the single word level natural language processing analysis shifts to the multi-word level, eventually, systems will be able to deal with complex "concepts" from multi-word expressions (Cambria and White 2014).

Conclusion

This research has presented a Sentiment-Semantic Framework to capture and use both sentiment and semantic analysis of online user-generated content. The framework is implemented as a prototype and tested on both online reviews and political discussion applications, from which the resulting insight knowledge bases are derived. A design science approach is taken to the development of the framework, its implementation, and assessment. The effective integration, use, and application of sentiment and semantic information can be considered a wicked problem, especially when dealing with unstructured, user-generated content (Rittel and Webber 1973). The insight knowledge base resulting from this research shows the feasibility, reusability, and comparability of combining sentiments and semantics. Future research is needed to expand the development of the prototype and its application as well as to apply it to other content such as blogs. Future research is also needed to generate large insight knowledge bases and test them for their usefulness in real world applications.

References

Abbasi, A., and Chen, H. 2008. "CyberGate: A design Framework and System for Text Analysis of Computer-Mediated Communication," *MIS Quarterly* (32:4), pp. 811-837.

Archak, N., Ghose, A., and Ipeirotis, P. G. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science*, (57:8), pp. 1485-1509.

Barsade, S., and Gibson, D. 2007. "Why Does Affect Matter in Organizations?," *The Academy of Management Perspectives ARCHIVE*, (21:1), pp. 36-59.

Baskerville, R., Kaul, M., and Storey, V.C. 2015. "In Design-Science Research: Justification and Evaluation of Knowledge Production," *MIS Quarterly*, (39:3), pp. 541-564.

Berners-Lee, T., Hendler, J., and Lassila, O. 2001. "The Semantic Web," *Scientific American*, (284:5), pp. 28-37.

Cambria, E., and White, B. 2014. "Jumping NLP Curves: a Review of Natural Language Processing Research." *Computational Intelligence Magazine*, IEEE, (9:2), pp. 48-57.

Chau, M., and Xu, J. 2012. "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities," *MIS Quarterly*, (36:4), pp. 1189-1216.

Chen, P. P. S. 1976. "The Entity-Relationship Model—Toward a Unified View of Data," *ACM Transactions on Database Systems (TODS)*, (1:1), pp. 9-36.

Clancey, W.J. 1993. "The Knowledge Level Reinterpreted: Modeling Socio-Technical Systems," *International journal of intelligent systems*, (8:1), pp. 33-49.

Daft, R. L., and Lengel, R. H. 1983. "Information Richness. A New Approach to Managerial Behavior and Organization Design," (No. TR-ONR-DG-02). Texas A&M University College Station, College of Business Administration.

Das, S. R., and Chen, M. Y. 2007. "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Management Science*, (53:9), 1375-1388.

Doan, A., Madhavan, J., Domingos, P., and Halevy, A. 2002. "Learning to Map Between Ontologies on the Semantic Web," in *Proceedings of the 11th international conference on World Wide Web*, ACM Press, pp. 662-673.

Fazzinga, B., Gianforme, G., Gottlob, G., and Lukasiewicz, T. 2011. "Semantic Web Search Based on Ontological Conjunctive Queries," *Web Semantics: Science, Services and Agents on the World Wide Web*, (9:4), pp. 453-473.

Frijda, N. H. 2007. "The Laws of Emotion," Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Garcia-Crespo, A., Colomo-Palacios, R., Gomez-Berbis, J.M., and Ruiz-Mezcua, B. 2010. "Semo: A Framework for Customer Social Networks Analysis Based on Semantics," *Journal of Information Technology*, (25:2), pp. 178-188.

Gruber, T. R. 1993. "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition*, (5:2), pp. 199-220.

Guarino, N., and Giaretta, P. 1995. "Ontologies and Knowledge Bases towards a Terminological Clarification," *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, IOS Press, Amsterdam, Netherlands, pp. 25-32.

Hendler, J., and Golbeck, J. 2008. "Metcalfe's Law, Web 2.0, and the Semantic Web," *Web Semantics: Science, Services and Agents on the World Wide Web*, (6:1), pp. 14-20.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly*, (28:1), pp. 75-105.

Liu, B., Hu, M., and Cheng, J. 2005. "Opinion Observer: Analyzing and Comparing Opinions on the Web," in *Proceedings of the 14th International Conference on World Wide Web*, ACM Press, pp. 342-351.

March, S. T., and Smith, G. F. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems*, (15:4), 251-266.

Metzler, D., and Croft, W. B. 2007. "Latent Concept Expansion Using Markov Random Fields," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, pp. 311-318.

Noy, N. F., and McGuinness, D. L. 2001. "Ontology Development 101: A Guide to Creating Your First Ontology," Knowledge Systems Laboratory, Stanford University.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, (2:1-2), pp. 1-135.

Qiu, G., Liu, B., Bu, J., and Chen, C. 2011. "Opinion Word Expansion and Target Extraction through Double Propagation," *Computational Linguistics*, (37:1), pp. 9-27.

Rittel, H. W., and Webber, M. M. 1973. "Dilemmas in a General Theory of Planning," *Policy Sciences*, (4:2), pp. 155-169.

Scherer, K. R. 2005. "What are Emotions? And How Can They be Measured?," *Social Science Information*, (44:4), pp. 695-729.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. 2003. "Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network," *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Association for Computational Linguistics, (1), pp. 173-180.

Turney, P., and Littman, M. L. 2003. "Unsupervised Learning of Semantic Orientation from a Hundred-Billion- Word Corpus," *ACM Transactions on Information Systems (TOIS)*, (21:4), pp. 315-346.

Ullmann, S. 1962. "Semantics: an Introduction to the Science of Meaning," Basil Blackwell, Oxford, United Kingdom, pp. 30.

Venable, J., Pries-Heje, J., and Baskerville, R. 2014. "FEDS: A Framework for Evaluation in Design Science Research," *European Journal of Information Systems*, advance online publication, November 11, 2014; doi:10.1057/ejis.2014.36.

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an Information System Design Theory for Vigilant EIS," *Information Systems Research*, (3:1), pp. 36-59.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 347-354.