Dec 11th, 12:00 AM

Information System Articulation Development - Managing Veracity Attributes and Quantifying Relationship with Readability of Textual Data

Neelam Naik
SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, neelam.naik@upgcm.ac.in

Shastri Nimmagadda
Curtin University, shastri.nimmagadda@curtin.edu.au

Seema Purohit
Affiliated with the University of Mumbai, supurohit@gmail.com

Torsten Reiners
Curtin University, t.reiners@curtin.edu.au

Dr Neel Mani
DSVV, neelmanidas@gmail.com

Information System Articulations - Managing Veracity Attributes and Quantifying Relationship with Readability of Textual Data

Completed Research Paper

Neelam Naik
SVKM Usha Pravin Gandhi College
Vile Parle, Mumbai, India
Neelam.naik@upgcm.ac.in

Shastri L Nimmagadda
Curtin University
Perth, WA, Australia
shastri.nimmagadda@curtin.edu.au

Seema Purohit
B. K. Birla College of Arts and Science
Kalyan, Mumbai, India
supurohit@gmail.com

Neel Mani
CAIR, DSVV
Uttarakhand, India
neel.mani@dsvv.ac.in

Torsten Reiners
Curtin University
Perth, WA, Australia
t.reiners@curtin.edu.au

Abstract

Often the textual data are either disorganized or misinterpreted because of unstructured Big Data in multiple dimensions. Managing readable textual alphanumeric data and its analytics is challenging. In spatial dimensions, the facts can be ambiguous and inconsistent, posing interpretation and new knowledge discovery challenges. The information can be wordy, erratic, and noisy. The research aims to assimilate the data characteristics through Information System (IS) artefacts that are appropriate to data analytics, especially in application domains that involve big data sources. Data heterogeneity and multidimensionality can make and preclude IS-guided veracity models in the data integration process, including customer analytics services. The veracity of big data thus can impact visualization and value, including knowledge enhancement in the vast amount of textual data qualitatively. The manner the veracity features construed in each schematic, semantic and syntactic attribute dimension in several IS artefacts and relevant documents can enhance the readability of textual data robustly.

Keywords: Big data, veracity attributes, Information system, readability, textual data

Introduction

The value of any data relies on uniqueness and quality, without making judgements on the validity and authenticity, including the readability of textual data. Various authors have interpreted Big Data in different characteristics, popularly in several Vs. Often the textual data are represented in various alphanumeric characters (Kiefer, 2019). The readability of textual data and their interpretability in various documents depend on the quality of words (appropriate vocabularies or text style), sentences, punctuations, grammar, and spell-checked expressions that describe legitimate conceptual and contextual attributes and instances. The information that describes textual data could be from multiple domains and systems, including diverse

concepts and contexts (Lacasta et al. 2010). We have proposed a holistic methodology to build relationships between veracity attributes and instances and ease the readability of textual data. We can ontologically structure textual facts to explore connections between wordy readable, erratic, and noisy data. For example, ontologically structured alphanumeric data can logically and physically be integrated into unified textual metadata in a repository system. The metadata can be cost-effective for explicitly exploring new knowledge in a textual form (Vasarhelyi et al. 2015).

The rest of the research paper is structured in various sections. The research objectives are highlighted, along with the motivation and significance of the research. The literature review identifies the research gaps in textual data mining. Various issues and challenges of Big textual Data mining are discussed in the next section. IS guided data mapping and modelling methods are described, including how the veracity attribute instances are mappable qualitatively with textual data mining artefacts that are construed with readability attributes and their occurrences, which are the highlights of the framework development. The value of information systems in various business contexts is highlighted in Rainer and Prince (2022). Several IS artefacts are articulated in the framework development. The integrated methodological framework adapts the ontology-based alphanumeric metadata with various mining artefacts, all accommodated in a repository. Results and discussions are analyzed, along with contributions and the intended research audience. The research is concluded with future scopes and limitations.

Literature Review

Big Data in information systems and their value chains, as research agendas with new opportunities and challenges, are reviewed with a profound interest in people, processes, and technology entities (Abbasi et al. 2016). The research agenda on Big Data analytics capabilities and their competitive advantage is discussed in Mikalef et al. (2018). For data science and Big Data analytics, statistical analysis is performed on veracity scores for spatial data (Chakraborty and Lahiri, 2019). The domains considered are wholesale trade, retail business, utility procurement, education, transportation, banking and securities, communication and media, manufacturing, government, and healthcare. The research is invaluable for Big Data system designers and future domain experts (Keskar et al. 2020). In this work, we study the relationship between the characteristics of Big Data and extract some categories from them. From this, we conclude that there are five categories, and these categories are related to each other (Gupta and Chaudhary, 2015). According to our research findings, there are no related measurements in information security management systems (ISO/IEC) for vital Big Data characteristics such as volume, variety, and variability. In future, theoretically, valid methods are planned to investigate for quality assessment of the Big Data V's (Omidbakhsh, M. and Ormandjieva, 2020). Many researchers have studied data veracity aspects and measurements, including timeliness, completeness, accuracy, and consistency. But none of the current methods provides a trustworthiness assessment of textual data generated over time. This paper introduces a model for fusing multiple data sources based on a multi-attribute decision-making (MADM) method with the consideration of uncertainty. In such contexts, a numerical example demonstrates the use of the attribute decision-making method (Amini and Chang, 2017). For example, text related to crisis intervention and suicide prevention organizations has saved lives using phone, instant messaging, and text massaging intervention services. With the increased use of social media and internet-based chat, many people post about their suicidal ideation on websites like Twitter and Facebook. Our initial results appear promising, and this research facilitates the usability of those analyzing social media text to improve the efforts of technology-based suicide prevention (Oseguera et al. 2017).

A comparative study is proposed between veracity models applicable to various networks, including Twitter and its facts (Al Doaies et al. 2017). The verification aspects considered during the comparison are working flow, accuracy, security, trust score, usability, and time consumption. Veracity attributed to irregularity, clutter and biases is frequently seen in Big Data (Howraa et al. 2020). The major challenge in analyzing Big Data is differentiating valuable information from profligate data. The data cleansing process is dedicated to enhancing the data quality by sensing and eliminating errors, contradictory material and overlaps and providing reliable, precise and consolidated data. In veracity analysis, it is observed that the data is relevant to the planned usage (Ylijoki et al. 2016). The two V's of Big Data, value and veracity, have not been the focus of characterizing Big Data but reflect their actual usage in textual data mining and analysis. However, 'veracity' is a critical V, gaining attention from the data analysis perspective because it is associated with the cleaning challenges of Big Data (Espinosa et al. 2019). Volume, velocity, and variety are three characteristics of Big Data. Veracity, a fourth V of Big Data, is gaining popularity because of problems building and maintaining extensive knowledge bases (Esteves et al. 2018). The issues are stated in the literature under the name's trustworthiness, fact-checking, and truth-finding. The text ascribes average sentence length, spelling mistakes, and abbreviations, which affect the quality of text mining results (Kiefer, 2019). Natural language

processing libraries such as Stanford Core-NLP and NLTK (Natural Language Toolkit) are used to measure quality indicators associated with textual data. The datasets from various domains, such as news, tweets, prose, and chat conversations, are used in the experiments.

Incompleteness is another attribute used in data analytics to estimate the text's noise level generated by social networks and Automatic Speech Recognition processing techniques (Gwenaelle and Lee, 2021). The incompleteness in Big Data refers to either noise in the labels or noise in the input data. The presence of noise in the data significantly impacts the prediction of any meaningful information. The classification accuracy decreases with noise in the data (Gupta and Gupta, 2019). The biggest challenge in developing a predictive model is to preprocess the data and make it noise-free. A systematic review of the existing literature is done, focusing on identifying and managing the text's noise patterns. In addition, various uncertainties associated with grey system data analytics are described in Yang (2019). The probability model is discussed to find objective uncertainty in small and big data. The principle of fact-checking is discussed by iteratively computing source reliability from the given dataset to calculate the veracity of the data through the reliability score of different sources (Berti-Equille, 2015). The Big Data veracity problem is presented through crowdsourcing solutions (Agarwal et al. 2016). Twitter data considered in the research use thousands of participants to probe and tag the tweets according to their sentiments. The sentiments are analyzed using ROC (receiver operating characteristic) curve and Bayesian predictor trained with trinomial function.

A recent survey suggests supervised and machine learning as two core procedures used for veracity assessment (Lozanoa et al. 2020). The research reiterates that the veracity assessment field is very complex, requiring a combination of data sources, document types, indicators, and methods to assess text's reliability precisely. Web events are considered a resource of Big Data by Wang et al. (2015). As a measure of veracity, an uncertainty estimation is planned by mining event features from the data of web cases. Linguistic feature-oriented readability formulas are examined through machine learning techniques based on the factors of individual readability (Liu et al. 2021). The readability measurement method is proposed based on a regression model that estimates text legibility. The technique supports linguistic characteristics such as discourse, lexical and syntactic notions (Kotani et al. 2011). The Flesch-Kincaid Grade Level (FKG), Flesch Reading Ease Formula (FRE), Gunning Fog Index (GF), Automated Readability Index (AR), Coleman-Liau Index (CL) and Linsear Write Formula (LW) are used for measuring the readability of the text (Zhang et al. 2019).

Issues and Challenges of Big Textual Data Mining

Big Data impacts in business intelligence and analytics are discussed by Chen et al. (2012). Though the benefits of Big Data are genuine and significant, many challenges remain to address and fully realize the potential of Big Data Analytics in text mining (Reihaneh et al. 2019). Some of the challenges are identifying the roles of characteristics of Big Data in text mining, existing analysis methods and models, and the limitations of the current word-based data processing system. The data processing and management of factual instances are crucial in Big textual Data analytics. The data challenges relate to characteristics such as volume, variety, velocity, veracity, volatility, quality, discovery, and dogmatism. The data process challenges are related to a series of techniques to capture, integrate, transform, and select the right model for analysis and provide the desired results. The management challenges cover privacy, security, governance, and ethics (Sivarajah, 2017).

One of the data challenges of Big Data analytics is the veracity of factual instances. The veracity feature measures the accuracy of data and its potential use for analysis (Vsarbelyi et al. 2015). In the current literature, veracity refers to coping with biases, doubts, imprecision, and fabrications with misplaced evidence in the data (Srivastava and Sahami, 2009). Veracity is measured to assess the complexity of textual data structure, anonymities, imprecision, or inconsistency in large data sets. It is not merely about data quality – but about understanding the data, as there are integral discrepancies in almost all the data collected. For example, every customer's opinion on social media networks is diverse and unclear, as it involves human interaction while managing textual data (Sivarajah et al. 2015). Dealing with inaccurate and ambiguous data in day-to-day transactions is challenging, as Gandomi and Haider (2015) discussed. Unfortunately, the significance of uncertainties in textual data analytics has not received sufficient attention that matches with challenges of big data-oriented data analytics research and applications (Yang, 2019). These research gaps have motivated us to explore new textual data modelling and mining.

What is Veracity in Textual Data?

The veracity attribute is a big data characteristic related to textual facts' consistency, accuracy, quality and

trustworthiness. In other words, veracity refers to noise, biases, abnormality, ambiguity, incompleteness and contradiction in the textual data. It is directly related to the transparency of data sources and the truthfulness of the data. The ability of the data to support a decision-making process is purposeful, useful, and of sufficient quality in the context in which it is analyzed (Crone, 2016). In Big Data analytics, textual data mining is robust and evolving tool. It enhances the power of unstructured textual data by finding new knowledge and significant patterns hidden in the data. The intricacy of numerous digital textual data has opened various research opportunities with new motivation and direction. The massive digital textual datasets are available as transcripts in the financial and marketing sectors, scientific literature, emails, social media posts, blog content, and web page content (Hassani et al. 2020). So, the current research focuses on modelling extensive textual data and analyzing schematic, semantic, and syntactic heterogeneities through various analytics tools and methods.

Why does Veracity of Big Data Matter in Information System Development?

Big textual data analytics is a complex process of examining large-size raw textual data to find hidden patterns, correlations, industry trends and customer preferences, including apprehensions in the end-user feedback and sentiments and other reasons that can be causative to misunderstandings of the information exchanged between various users. Misinterpretations of the textual data can be barriers to future business analyst's aspirations. During contextual application development, motivated in textual data analytics, information systems permit end users to collect, store, organize, distribute, and analyze volumes and varieties of textual data, including their meaningful interpretations in various business activities and functions. The applications can serve multiple users to achieve commercial goals and objectives. In addition, the process can help organizations make improved decision-making in alignment with current business and market needs. So, accuracy and precision in textual data quality can help improve information system articulations with better-informed business decisions.

Why is Veracity Calculation Important?

The term veracity refers to the quality of the data in terms of uncertainty and impreciseness. According to the estimation of IBM, due to the poor data quality, the cost bared by the US economy is around $3.1 trillion per year. The veracity of the data is categorized as good, bad, and undefined based on attributes of data such as incomplete, noisy, ambiguous, and inconsistent. As diverse and varied data sources are involved in data generation, the accuracy and trust of the data need to be evaluated before performing the analytics and interpretation of the data. To give an example,

- People prefer social media platforms like Twitter to share official corporate information, personal views and opinions. However, confusion may arise with difficulty in analyzing Twitter data using current techniques and limited sentiments of people.
- Another example is related to the healthcare system. The disease trend or outbreak can be identified by analyzing billions of healthcare records. But the inconsistencies and ambiguities in such a dataset may result in the false prediction about the diseases, with result, an increase in outbreak is observed (Reihaneb et al. 2019).

Quantification of Veracity Attributes

To measure the overall veracity of Big Data (AlDoaies et al. 2017), we need to quantify the reliability of attributes in different parameters, scales, and magnitudes. The data relating to quantification parameters considered in the present study are noise, biases, ambiguity, abnormality, and incompleteness.

Noisy textual data – The irrelevant information in the data is called noise. The data that is not understandable to the user system and impossible for the user to interpret such facts is called noise. The noise in the text distorts the content, including semantics and syntactic information, implying that noise in the data can adversely affect textual data analytics. Inconsistencies between electronic documents, the original textual details, and the variations may be due to noise. Two primary sources of noisy text can mislead factual interpretation. Noise can appear in the text when multiple sources capture textual information. For example, various sources of information are either printed or hand-written documents or camera-captured images of the document. The noise can enter the final text during conversion because of the wrong interpretation of words or letters. Even noise is introduced at the time of production of the digital text. Typical examples of such digital text are Short Messaging Service (SMS), emails, web page content, text generated by Automatic Speech Recognition (ASR) system, text generated by Optical Character Recognition (OCR) device and live chats. The noise in such sources is in the form of grammar mistakes, use of multi-lingual words, special characters,

pause-filling words (such as um, uh), repetitions of the word, abbreviations, and non-standard word forms. Thus, the noise can affect text details during the conversion and generation of the digital text document.

The widespread error occurring in the text is a spelling mistake. One of the standard methods to measure spelling mistakes is edit distance. If there are two strings, s1 and s2, the edit distance between them is the minimum number of edit operations required to transform s1 into s2. The edit distance between CAT and RED is 3, while the edit distance between CAT and RAT is 1. Edit operation includes deletion, insertion, or replacement of a character in a string. Different weights are assigned to different edit operations. For example, replacing character *s* with *a* has a higher weight than character *s* with a character *p, which is* incomparable to *s* on the character typing keyboard. In short messaging services (SMS), non-standard spellings, abbreviations, and phonetic transliteration are the few sources of noise in the text. The particulars of SMS noise are as follows.

- Character deletion: The commonly identified pattern is 'msg' for the word 'message' or tom for the word tomorrow
- Phonetic substitution: The insertion of digit 2 for the word 'to' or 'too'.
- Abbreviation: The frequently used abbreviations are 'lol' for 'laugh out loud'.
- Dialectal and informal usage: The multiple words are combined into a single token. The example is 'gonna' for 'going to', 'aint' for 'are not'.
- Deletion of words: The function words and pronouns are typically deleted. 'I am driving back home' is typed as 'drving hm'.

The Word Error Rate (WER) is the metric used to measure noise level in SMS data at the word level.

$$WER = \frac{S + D + I}{N}$$

S is the number of word substitutions, D is the number of word deletions, I is the number of word insertions, and N is the total number of words. The Sentence Error Rate (SER) is the metric used to measure the noise level in SMS data at the sentence level.

SS is the number of sentence substitutions, DD is the number of sentence deletions, II is the number of

$$\text{SER} = (\text{SS} + \text{DD} + \text{II})/\text{NN}$$

sentence insertions, and NN is the total number of sentences (Subramaniam et al. 2009). The incorrect data instances can make contradictory textual facts. The application of corrections in the spelling of words is an example of contradiction removal from the text content. Also, applying the algorithm to correct sentences grammatically and syntactically is part of contradiction removal. The present study focuses on the two attributes of veracity considered within the judgement of attributable noise. The sub-attributes are percentages of spelling and grammatical mistakes in each text.

Ambiguity in semantics and syntactic text – the data quality can make the same text interpreted in several ways, which cannot make sense, which is called ambiguity in the data. Ambiguity is an intrinsic property of natural language. The capability of being understood or interpreted in two or more ways is considered an ambiguity. In textual data processing and analysis, the application's critical part is identifying ambiguous words, phrases, and sentences. The second factor associated with ambiguity is uncertainty, implying a lack of confidence about the meaning because of a gap in the writer's and the reader's background knowledge or both. Traditionally, ambiguities have four types: lexical, syntactic, semantic, and pragmatic. When a single word has many meanings, then lexical ambiguity occurs. Homonymy lexical ambiguity can arise when two words have the exact written and phonetic representation but different histories during textual data development. Polysemy lexical ambiguity occurs when a phrase has one etymology and many related meanings. Syntactical ambiguity arises when the sequence of words can have more than one grammatical structure, but each one has a different meaning with different syntactic rules. Even though there is no lexical or structural ambiguity, and if the sentence has more than one way of reading or interpreting it within its context, such ambiguity is called semantic ambiguity. Pragmatic ambiguity occurs when a sentence has multiple meanings in the context. The present study focuses on the lexical ambiguity associated with words (Kiyavitskaya, et al. 2008). An ambiguity in the text measures the number of words and phrases having multiple meanings. The present study concentrates on measuring the ambiguity of the text in terms of the percentage of words having many meanings.

Abnormality – The data that falls outside the acceptable, readable, or interpretable range is called abnormal data. The present study measures the irregularity of the given textual dataset in terms of the

percentage of an unknown word and the rate of short forms and abbreviations in each given textual dataset.

Incompleteness – The data is said to be incomplete if it has missing values of the attributes representing the factual text. The short words and sentences contribute to the incompleteness attribute of the overall textual data. The present study measures the incompleteness in terms of incomplete words in the given text dataset.

Biases – If the available data is not representative of the phenomenon under study, then it is said that the data have bias. The given textual data may be missing the variables to capture the phenomenon correctly. Opinion and sentiment classification are two approaches of text mining mainly preferred to remove bias from the textual data. The main aim of opinion classification is to determine to what extent the text under consideration supports or opposes the subject. Opinion classification calculates the text polarity and decides whether a given text is biased or unbiased depending on the text with a 'positive' or 'negative' opinion on the subject matter. An unbiased text refers to the comments on its subject matter and expresses a specific opinion on the context. Another technique of bias removal from textual data is sentiment classification. It is closely related to opinion classification. Based on polarity, tone, valence and appraisal, sentiment classification extracts effective content from the text. It uses pre-defined lists of terms having already allotted quantitative weights for positive and negative connotations. Based on these connotations, the sentiment score is calculated.

$$Sentiment\ Score = \frac{No.of\ positive\ terms - No.of\ negative\ terms}{No.of\ all\ terms}$$

If the weights already exist, then consider the importance,

$$Sentiment\ Score = \frac{\sum_j W_j^+ - \sum_k W_k^+}{\sum_j W_j^+ - \sum_k W_k^+}$$

Where $W_j^+$is the sentiment weight for j[th] positive term and W_j^- is the sentiment weight for j[th] negative term (Hassani et al. 2020).

Readability of Text

The accuracy of the given text is measured in terms of readability. The readability of the text provides feedback to the writer about how effectively he or she reaches the audience. Various readability score determination methods and formulae include Dale–Chall formula, Gunning fog formula, Fry readability graph, McLaughlin's SMOG formula, FORCAST formula and Flesch Scores method. In the current experiment, the Flesch Scores method is used to measure the readability of the text. According to this method, the "reading ease" score is calculated with the help of ASW and ASL:

$$\text{Reading Ease score} = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

ASL = average sentence length (the number of words divided by the number of sentences), and ASW = average word length in syllables (the number of syllables divided by the number of words).

Motivation and Significance of Research

The existing readability and veracity challenges have motivated us to explore new textual data models and their implementation in various applications. The significance of the research lies in the facts of the correctness and trustworthiness of Big textual Data sources and exploring their veracity (Al Doaies et al. 2017). Veracity models have been compared in social media platforms to assess the richness of the volume of Big Data. Several datasets considered in the modelling process test the veracity attributes and can read their instances accurately. For example, the data attributes pertained to customer complaints, issues with ABC (Australian Broadcasting Corporation) news, coronavirus tweets, fake news, travel advisories, women's clothing, child book and short stories, and their instances are carefully reflected in the modelling process. Using appropriate vowels and consonants has motivated us to articulate the words and sentences without any errors and mistakes and further validate the readability of words and sentences.

Research Questions and Objectives

Logically organized words and their alphanumeric character settings can assess the document's quality and the user's readability. The documents' transcripts may comprise multiple conceptualized and contextualized semantic, schematic, and syntactic expressions (Alemi and Ginsparg, 2015). Ontological relationships between various words and alphanumeric characteristics of Big textual Data can facilitate improvements in the veracity and quality of documents with readability judgements (Rudra and Nimmagadda, 2005 and Lacasta et al. 2010). The research aims to establish the relationship between the various attributes of veracity and the reading accuracy of the text. A couple of research questions are designed (1) how the veracity attributes of Big Data affected textual data mining? (2) How do we evaluate the veracity models, assessing the reliability of mining models and facilitating the readability? For the readability of the textual data, we have aimed a couple of research objectives (1) design and develop IS artefacts with integrated methodologies that can unify various alphanumeric characters and assimilate the readability of textual data, (2) Evaluation of IS and mining models that can clarify the readability of textual data instances. The evaluation includes the textual data's schematic, semantic and syntactic heterogeneity assessments.

IS Modelling Methodology for Managing Veracity of Textual Data

Opportunities of Big Data, Data Science and Analytics and challenges in IS research are discussed by Agarwal and Dhar (2014) and Wu et al. (2014). Based on the research gaps in the existing literature, we have focused on improving consistency, accuracy, quality, and trustworthiness, including readability and reproducibility, through various mapping and modelling techniques. Data veracity refers to bias, noise, and abnormality in data. The sentences can be incomplete with errors, and outliers, including missing instances. To assess the veracity of data, we must ascertain the original data source, contributed bases, and periodic dimensions of the textual data, including the methodology of data acquisition from reliable sources.

Objectivity, truthfulness, and credibility are other dimensions usable in the dimensional modelling process. The mapping and modelling process evaluates the power of conveying the semantics to the readers. Building relationships between various attribute dimensions can make veracity with meaningful research objectives and goals of textual data readability. Characteristically, we have used vowels and consonants that offer quality wordy sentences and documents to review further the noisy and erratic data sources (Block and Duke, 2015 and Nimmagadda et al. 2019). The logic behind the use of vowels and consonants in constructing meaningful words and sentences is discussed by Block and Duke (2015).

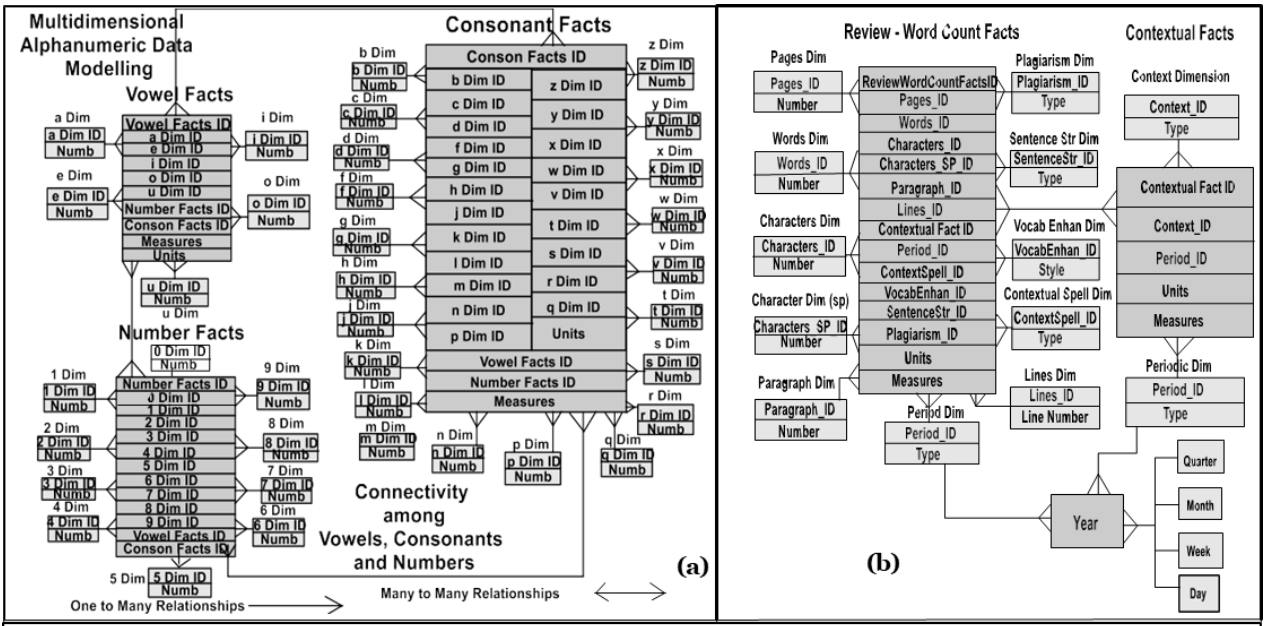

Figure 1. Information System (IS) Artefacts Articulated to connect the Alphanumeric Characters, Letters and Words (a) Modelling Alphanumeric Characters (b) Schema with Word Counts and Contextual Facts

Figures 1a and 1b demonstrate several sequential vowels and consonants, interpreted as attribute dimensions in dimensional modelling (Nimmagadda et al. 2019). For example, the vowels “a” to “u” and “b” to “z”, including “0” to “9”, are interpreted as dimensions in the dimensional modelling process, as shown in Figure

1. The purpose of the schematic structures is to interconnect the alphanumeric characteristics in diverse contextual textual documents. Data integrity can be pragmatic during repository designs, including model-building process operations. Logical and physical integrities that refer to primary data types are made functional in textual data analytics, which can be practical in building repository systems and involving big textual data sources. Cloud-based service companies offer objectively manageable veracity of textual data analytics and business intelligence. Logical integrity facilitates to protect from human errors, including malicious intervention. Other forms of logical integrities are entity integrity, referential integrity, domain integrity and user integrity. Vowels and consonants can be entities or dimensions in the modelling, logically interpreted as primary and secondary keys that link the vowels and consonants in various ways that make schematic and semantic sense. Textual data access and storage must ensure the referential integrity that can associate various tables within a relational database structure. Data about multiple domains and document ecosystems must be ascertained to acceptable and readable levels, ensuring textual data instances are interpretable without ambiguity or bias. User integrity is a key challenge in ensuring business rules and constraints are imposed on the data structures so the user can use them for a specific purpose. In the current context, readability is a user-specific evaluation in addition to the quantification of veracity attributes.

As per research objective 1, we have articulated a framework to address the connectivity between words and sentences in the textual data. These articulations can motivate us to minimize human errors and mistakes during the construction of sentences with appropriate words and semantics that make sense. We thus develop a framework to accommodate textual data structures that describe different types of vowels and consonants in words and sentences. Ontologies are interpreted wherever necessary to interconnect these vowels and consonants to make and construe suitable words and sentences in various interpretable contexts. Transforming unstructured textual data into meaningful information demands new integrated frameworks, which can be applied in various contextual applications (Alwidian et al. 2015). As demonstrated in Figure 2, we articulated a framework with various tools and utilities in the form of IS artefacts to integrate textual data from multiple sources. Various objects are used to construct IS artefacts, as demonstrated in Figure 2. Each object performs a particular task, such as grammar or spell-checking. Ontologies are construed in interconnecting various sub-schemas. Various tools use the framework development, such as domain/context checker, thesaurus, word checker and validator.

Several frameworks are available to effectuate improved readability of text (Sowmya, 2021). But in the current research, we articulate a framework, construing domain ontologies with taxonomic class hierarchies, class definitions and class conceptualization of words and alphanumeric characters that represent multiple dimensions. In addition, business rules and axiom constraints are imposed on the modelling process. The purpose of the framework is to unify, assimilate and assess the complex relationships between words that represent multiple attributes of Big Data articulated within a repository system. Thesaurus, word validator and checker, including grammar checker, are connectable to ontology structures to make words and their associated documents semantic, schematic, and syntactic. The method of quantifying the veracity attributes is detailed in Table 1. Noise, ambiguous, incomplete, and abnormal data are typical veracity attributes interpreted to assess the articulated framework in Figure 2. For example, we have quantified the veracity attributes relevant to finding miss-spelt words, grammar mistakes, words with multiple meanings, and abbreviations, including unknown/uncommon words, all through different syntactic views. We have examined for missing data and their inaccuracy that cannot make sense of semantics and any new insights for interpretation. Data mining and visualization techniques assess measurable levels of trust, consistency, accuracy, quality, and trustworthiness (Nimmagadda et al. 2019). Figure 2 is a framework that can holistically assess the veracity characteristics of Big Data. For readability, several stages are initially described, such as spell check before and after modelling, verifying the words by validator, domain/context checker, and validating the grammar by thesaurus. In addition, we have measured veracity attributes that represent noisy, abnormal, and incomplete words. Table 1 describes veracity attributes and the corresponding quantification method of those attributes that can make out ambiguous words.

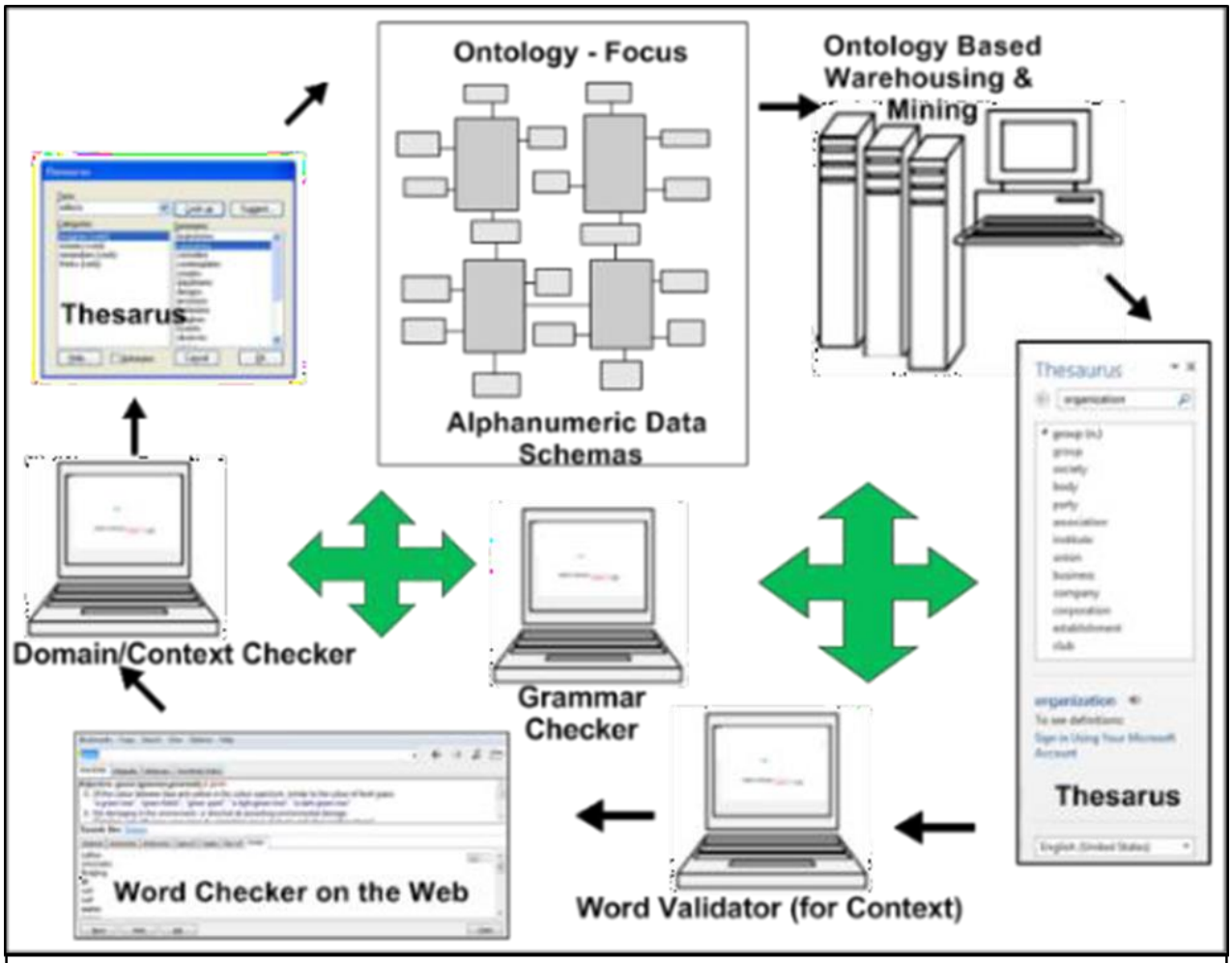

Figure 2. Framework Development that Manages Multidimensional Alphanumeric Characters and their Readability

Veracity attribute	Method of Quantification
Noise	Percentage of spelling mistakes
Noise	Percentage of grammatical mistakes
Ambiguity	Percentage of words having many meanings
Incompleteness	Percentage of incomplete words
Abnormality	Percentage of unknown words
Abnormality	Percentage of short-forms
Abnormality	Percentage of abbreviations

Table 1: Method of quantification for veracity attribute

A grammar-guided genetic programming algorithm for associative classification in Big Data is provided by Padillo et al. (2019). Currently, Python code is designed to read the text file line by line. Each word of each sentence is parsed. Many Python libraries are used to count spelling mistakes, grammatical mistakes, words with many meanings, unknown or uncommon words, short forms, and abbreviations in the text content. Some Python programs written to find the percentages explained in Table 1 are shown in Figures 3-7.

```
# counts spelling mistakes
from autocorrect import Speller
spell_check = Speller(lang='en')
if word != spell_check(word):
incorrect_spell_count= incorrect_spell_count+1
```

Figure 3: Python Code to Find Miss-spelled Words

Python autocorrect package is installed for finding spelling mistakes in a given text. From this package, the Speller class is imported. This class's object is initialized with lang='en' as an argument. Figure 3 shows the code snippet for finding the spelling mistakes.

```
#counts number of grammatical
errors. #percentage of grammatical
mistakes import nltk.data
nltk.download('punkt')
import language_tool_python

tool = language_tool_python.LanguageTool('en-US')
original_sen = '\n'.join(sent_detector.tokenize(text.strip()))
print(original_sen)

correct_sen = tool.correct(original_sen)
print(correct_sen)


gramatical_errors = 0


for line,line1 in zip(original_sen,correct_sen):
   if line!=line1:

      gramatical_errors = gramatical_errors +1


print ('Number of grammatical mistakes = ',gramatical_errors)
print ('Percentage of grammatical mistakes = ',gramatical_errors*100/len(original_sen))
```

Figure 4: Python code to find grammatical mistakes

Python nltk.tokenize package contains a tokenizer class that divides a string into substrings by splitting on the specified string, defined in subclasses. Figure 4 shows the code snippet for finding the grammatical mistakes.

```
#counts multiple meaning words

from PyDictionary import PyDictionary
list_of_acronyms_of_word = dic.meaning(word)
if type(list_of_acronyms_of_word) is dict:
multiple_meaning_words = multiple_meaning_words+1
```

Figure 5: Python code to find multiple meanings words

Python package PyDictionary is used to find words having multiple meanings. The various meanings of the word are stored in the list data structure. Figure 5 shows the code snippet for finding the words having multiple meanings.

```
import re
#counts total number of abbrevations
abb_text = re.sub(r"\b[A-Z]{2,}\b", "", word)
if abb_text=="":
      print('abbrevation=',word)
      abbrevations_count = abbrevations_count+1
```

Figure 6: Python code to find abbreviations

Python regular expression library re is used to find abbreviations in a given text. Figure 6 shows the code snippet for finding the words in abbreviations or short-form format.

```
#counts uncommon words in english language from a given text file
import enchant
d = enchant.Dict("en_US")
if not d.check(word):
uncommon_word = uncommon_word + 1
```

Figure 7: Python code to find unknown/uncommon words

pyEnchant spellchecking python library finds unknown and uncommon words from the given text. Figure 7 shows the code snippet for finding the unknown and uncommon words in each text content.

Relationship between Veracity Attributes and Readability of Text

The text-based ten datasets are considered in the experiment. These datasets are downloaded from the Kaggle website. These datasets are purely text-based (www.kaggle.com/datasets). A relationship between veracity and readability is interpreted using textual data models and the integrated framework designed in Figure 2. Metadata is computed, and various data views are extracted from the integrated framework and deduced to assess textual data's readability. Simple coding snippets are used (Figures 3-7) to compute interpretable data views.

Datasets Considered

cust_complaint is the dataset about consumer complaints on financial products. The consumer complaints are about debt collection, consumer loans, money transfers, student loans, bank accounts, credit reporting and credit cards. constitu_India is the dataset of the Constitution of India. This dataset contains articles from 1 to 395 and their sub-articles also. child_book dataset is related to highly-rated children's books. It includes information on the target age group for which the book is written and the corresponding book's description. women_clothing dataset contains reviews from women customers on the E-commerce clothing website about purchased products. EA_shortstories contains 69 Edgar Allan Poe's short stories in a tabular format having columns to describe publication date and classification made by Wikipedia. trip_advisor is the dataset of about 20000 hotel reviews extracted from Trip Advisor. fake_news dataset contains a list of articles that are considered false news. NY_times_comments dataset includes information on people's opinions and current affairs. corona_tweets dataset contains the tweet of users hash tagged with coronavirus, coronavirusoutbreak, coronavirusPandemic, covid19, ihavecorona, and stayhomestaysafe. The dataset abc_news contains around a million news headlines published for 18 years. The dataset has two columns, the date of publishing and the headline text in ASCII English lowercase.

Results and Discussions

Qualitative analysis is carried out on selected textual data using various software tools. The completeness of a document hinges on readable textual fact instances, filled with several pages, different paragraphs, sentences, words, and flawless alphanumeric characters. The textual data is, however, dependent on attribute variables that can be interpreted during validation of the veracity and readability instances. The linear regression model determines the character and strength of the association between a dependent variable and a series of other independent variables, either in words or several sentences. In other words, modelling study suggests that textual data can still exhibit syntactic, schematic, and semantic heterogeneities. The data instances used for executing the Python codes on various datasets are shown in Table 2.

Dataset	spellMis	graErr	wordsMulti	incomWords	unknWords	abbrev	avgSenL	readability
cust_complaint	94.85	84.61	86.35	19.63	46.39	14.5	82.56	66.74
constitu_India	10.57	10.23	45.56	0.5	45.26	8.96	74.56	78.26
child_book	14.56	15.23	50.41	16.94	12.56	5.63	56.24	85.22
women_clothing	81.42	85.42	45.66	56.49	67.56	45.52	64.21	45.97
EA_shortstories	12.32	16.28	40.12	0.8	10.56	8.52	40.25	96.56
trip_advisor	16.59	11.45	55.63	48.52	48.67	9.63	57.86	78.54
fake_news	99.45	78.94	87.2	89.69	88.74	78.98	59.47	49.63
NY_times_comments	17.89	15.46	69.87	14.65	18.47	56.94	81.94	82.56
corona_tweets	14.56	11.23	57.68	85.69	57.22	7.96	59.45	77.52
abc_news	16.56	17.61	62.23	12.36	15.46	49.56	86.36	84.27

Table 2: Veracity attribute instances for various datasets

The linear regression algorithm is applied to Table 1 using Weka 3.8.1. Internally the data is normalized, and the linear regression algorithm is applied. The following equation I gives the model generated.

Readability = - (0.2 * spellMis) – (0.2735 * graErr) + (0.306 * wordsMulti) – (0.0495 * incomWords) – (0.2954 * unknWords) – (0.1916 * abbrev) – (0.0264 * avgSenL) + 0.7626 .....(I)

The Sequential Minimal Optimization algorithm is used for linear regression.

D. No	INCS	MSP	CPE	G	P	SS	S	VD	P	RS (%)
1	1	3	1	1	1	0	1	2	1	97
2	2	1	0	0	0	0	0	2	0	100
3	3	3	1	1	1	1	0	4	2	98
4	6	4	2	2	2	2	1	4	2	95
5	5	6	3	2	3	2	2	3	3	97
6	5	7	4	4	4	2	2	1	3	93
7	4	3	5	3	2	1	2	1	2	97
8	2	3	3	1	1	0	0	1	0	100
9	2	3	2	2	1	1	1	4	2	98
10	3	5	4	2	2	2	2	4	2	95
11	3	6	5	2	4	2	2	5	3	97
12	4	5	6	3	5	4	3	3	4	93
13	3	5	5	3	4	3	4	2	4	95
14	2	3	6	3	4	4	3	3	2	97
15	3	4	6	4	4	5	4	4	2	97

Table 3: Number of documents analyzed for veracity attributes and instances

D. No: Document number; INCS: Incomplete sentences; MSP: Misspelled; CPE: contextual spelling errors; G: grammar predicaments; P: questionable punctuations; SS: sentence structure issues; S: style; VD: vocabulary debility; P: plagiarism; RS: readability score (%).

In addition, we have analyzed several multi-linguistic documents for veracity and readability attributes to examine schematically, semantically, and syntactically with a purpose and establish the relationship between

variability features as detailed in Tables 3 and 4. Evaluation of veracity attributes is done through a framework, a common platform where various IS artefacts attributable to veracity characteristics and their connectivity phenomena are examined, as in Figure 2, through which missing links between textual data are explored and assessed. We further detail, as discussed in schemas in Figures 1a and 1b, each word is checked for appropriate vowels and consonants, including several alphanumeric characters and word counts. These models are IS artefacts accommodated in a framework theory in Figure 2. The theory is further theorized with algorithmic Python codes in Figures 3-7. IS approach additionally demands an instantiation process to work and make the application in the current contexts work. For this purpose, we examined several documents to evaluate veracity attributes and readability contexts (Nimmagadda et al. 2019). For example, attributes such as syntactic rules, similarity, and dissimilarity in semantic words, including grammatically correct phrases and semantically correct words that make sense in the meaning, syntactic structures and identifying relationships between individual words in particular contexts and their instances are used in the analysis for 15 different types of documents. Various attributes are assessed, as shown in Tables 3 and 4.

DN	NW	NCS	ANSY	NR	SS	GCW	DSS	DSYNS	SCW	IRBIWC	RS (%)
1	4	4	5	1	1	1	0	1	2	2	97
2	7	5	6	1	0	4	0	1	3	6	100
3	7	5	6	3	2	5	0	1	3	7	98
4	8	7	5	4	3	4	2	2	2	6	97
5	8	6	4	3	3	3	2	3	2	5	97
6	4	4	3	2	2	1	2	1	1	5	93
7	5	5	4	2	1	2	1	2	2	5	97
8	7	6	8	2	2	5	1	3	3	8	100
9	5	7	5	4	3	3	2	2	2	6	98
10	5	4	4	2	2	2	1	1	1	5	95
11	6	7	6	3	4	3	2	3	2	6	97
12	5	5	5	2	2	2	2	2	1	5	93
13	5	6	5	2	3	3	1	2	3	6	95
14	4	5	5	3	2	2	1	3	2	5	97
15	4	3	3	2	2	1	1	1	1	2	93

Table 4: Number of documents analyzed for schematic, semantic and syntactic readability

D: Document number; NW: Number of words in sentence; NCS: Number of contexts in a sentence; ANSY: Average number of syllables per word per sentence; NR: Number of rules; SS: Similarity in semantics: GCW: Grammatically correct words; DSS: Different semantic structures; DSYNS: Different syntactic structures; SCW: semantically correct words; IRBIWC: Identifying relationships between individual words and contexts; RS: Readability Score (%)

As demonstrated in Figure 8a, we have analyzed various veracity attributes relevant to readability. A couple of unreadable trends are interpreted using attributes of incomplete sentences, miss-spelled, contextual spelling errors, and grammar and punctuation issues, including vocabulary debility and plagiarism challenges. For examining the schematic, semantic and syntactic readability scores, we have considered various veracity attributes that are interpretable in multiple documents (Martinez-Gil, 2023).

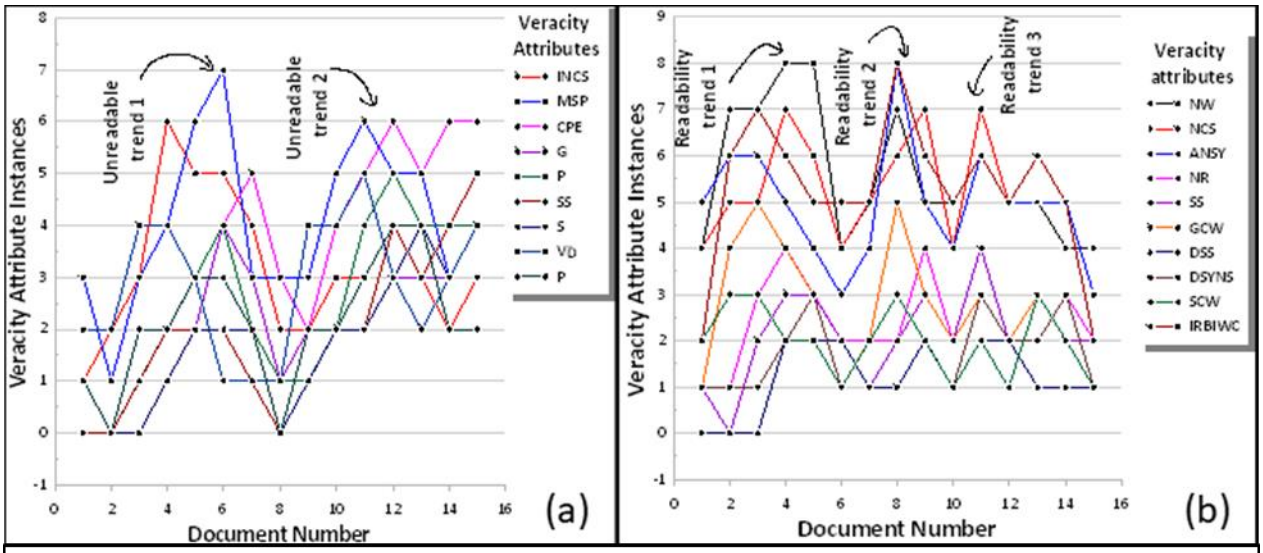

Figure 8. Line/scatter plot views – (a) Unreadable Trends interpreted from the Document Ecosystems (legend is described in Table 3); (b) Veracity Attribute Instances, showing the Readability Trends for Different Document Ecosystems (the legend is described in Table 4)

Attributes such as the number of words in a sentence, the number of contexts in a sentence, the average number of syllables per word per sentence and the number of rules are examined for readability. Three typical readability trends are interpreted as shown in Figure 8b, where high veracity attribute instances are reported. Similarly, semantic, grammatically correct phrases, possible semantic structures, including probable syntactic structures, semantically correct words and relationships identified between individual words and contexts are analyzed, sure to improve readability trends, as demonstrated in Figure 8b.

Probable Research Applications

Transforming unstructured textual data into meaningful information demands new integrated frameworks, which can be applied in various contextual applications. We have examined volumes and varieties of textual data to find correlations, patterns and trends and interpret new insights into how textual veracity data can play roles in various business applications. The research has scope and opportunity in a variety of business applications. A few veracity-based textual data mining applications are summarized here:

Customer service: Veracity has a role in detailing customer textual data, including surveying customers' opinions and translating the feedback from Twitter textual data and through chatbot systems. Customer satisfaction and better communication in real-time between customers and various clients of multinational companies can facilitate businesses without any ambiguities. As elaborated in the Results and Discussions Section, the methodological framework discussed in Figure 2 can enable the customer data to connect the veracity attributes with readability attribute instances.

Financial risk management: Extracting useful information from textual data analytics has significance at a workplace for reporting purposes. Qualitative analysis of veracity attribute patterns and trends, as discussed in Figures 8a and 8b, can help understand the information on business trends and performances of financial institutions rapidly to make accurate decisions, besides minimizing the risk of financial management.

Business maintenance: Maintenance of business depends on functional and non-functional requirements and business goals. Text analytics can help assess the functional requirements in businesses that offer quality products and services to various customers, including decision-making through methodologies discussed in the research. Non-functional requirements such as the performance of mining tools and new knowledge interpreted from business textual data contexts and applications can help maintain the overall businesses with achievable goals and customer sentiments.

Healthcare analysis: Manual investigation of sensitive healthcare data is time-consuming to make medicinal decisions. The current research has relevance in many industries reducing the turnaround time of analyzing textual data, especially the qualitative analysis of healthcare textual data, based on which useful

information can be extracted. The correlations, trends and patterns interpreted from text mining based on which anomalies observed in the textual data analysis can identify outliers in the healthcare data.

Spam filtering through textual analytics: Textual data mining can identify spam and hackers' activity, which can infect the computing systems or even gain access to sensitive customer data and information illegally.

These methodologies may save enormous computational power and time if the procedure is explicitly and judiciously applied, including with adaptable and implementable iterative IS artefacts. The data quality is crucial; if the quality is well maintained, the data mining, visualization and interpretation artefacts of the framework may improve the veracity and readability scores. In terms of benefit-cost ratio analysis, the cost of the study is minimal compared to the project's overall benefit in qualitative terms.

Research Audience and Contribution

Researchers involved in text mining and qualitative and quantitative analysis of veracity attributes and text mining are the beneficiaries of the study. Natural language processing personnel who can interpret, manipulate, and comprehend the human language are other audiences who can contribute to building relationships between veracity attributes attributable to document ecosystems. Constraints may be associated with data acquisition methodologies and the data sources where the reliability may be evaluable. More research is needed in testing and establishing the relationships between veracity and readability score attributes in multiple application domains.

Conclusions and Limitations

Textual data in multiple documents are categorized as Big Data, mainly representing volumes, varieties, and veracity of alphanumeric characters with various dimensions. The IS artefacts and the integrated framework developed have significance and robustly analyze the veracity of textual data. IS artefacts can interconnect the veracity attributes construed as entities, dimensions and objects in the modelling process. In this research, several veracity attributes are analyzed. For text mining and analytics, we analyzed fifteen documents in total. Snippets of Python code helped us extract veracity attributes from textual data. From the experiment, the percentage of spelling mistakes, grammar mistakes, incomplete words, unknown words, abbreviations, and average sentence length is negatively correlatable with the readability of the text. Readability scores rely on the quantification of veracity attributes. The percentage of multiple-meaning words positively affects the readability of the text, implying that while reading the textual content, multiple- meaning words can easily be grasped concerning the contexts by the reader. However, the quality of veracity attribute instances has constrained the document quality because of the textual data's complexity and variability. However, it is recommended to consider several documents for veracity attribute assessments. Data acquisition methodologies and the reliability of data sources must be ascertained before veracity attributes are assessed. Data quality is another challenge which can be resolved by data mining, visualization, and interpretation techniques. More research is needed in testing and establishing the relationships between veracity and readability score attributes in multiple application domains.

Future Work

We plan to carry out and robustly incorporate systematic schematic, semantic and syntactic readability attributes and scores that can be integral to the entire legibility and veracity of the textual data. In addition, we intend to apply the methodologies in different applications.

References

Abbasi, A. S., Suprateek & Chiang, R. H.L. (2016). "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems*, 17(2), DOI: 10.17705/1jais.00423.

Alwidian, S., Bani-Salameh, H. & Alaa, A. (2015). Text data mining: a proposed framework and future perspectives. *International Journal of Business Information Systems*. 18. 127-140. 10.1504/IJBIS.2015.067261.

Agarwal, R. & Dhar, V. (2014). Editorial-Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS research. *Info. Sys. Res* 25 (3): 443-448. https://doi.org/10.1287/isre.2014.0546.

Alemi, A. A. & Ginsparg, P. 2015. Text Segmentation based on Semantic Word Embeddings, KDD '15 Sydney, Australia.

Amini, M., & Chang, S. (2017). Assessing data veracity for data-rich manufacturing. In *IIE Annual*

Conference. *Proceedings* (pp. 1661-1666). Institute of Industrial and Systems Engineers (IISE).

Al Doaies, B. H., Ashi, A. M. & Alotaibi, F. S. (2017). "Exploring and Evaluating the Impact of the Veracity of Big Data Sources", International Journal of Computer and Information Technology (ISSN: 2279 – 0764), Volume 06 – Issue 05, September 2017.

Agarwal, B., Ravikumar, A. & Saha, S. (2016). "A Novel Approach to Big Data Veracity using Crowdsourcing Techniques and Bayesian Predictors", *ACM COMPUTE* '16, October 21-23, 2016, Gandhinagar, India, 2016 ACM. ISBN 978-1-4503-4808-9/16/10, DOI: 10.1145/2998476.2998498.

Berti-Equille, L. (2015). "Data Veracity Estimation with Ensembling Truth Discovery Methods", 2015 IEEE, *International Conference on Big Data (Big Data*).

Block, M. K. & Duke, N. K. (2015). Letter Names can cause confusion and other things to know about letter-sound relationships, Young Children**,** March 2015**,** Vol. 70, No. 1, National Association for the Education of Young Children.

Chakraborty, A. & Soumendra, L. (2019). On Statistical Properties of a Veracity Scoring Method for Spatial Data. ARxIV:1906.08843, https://doi.org/10.48550/arXiv.1906.08843.

Chen, H., Chiang, R. H. L. & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact**,** *MIS Quarterly*, Vol. 36, No. 4 (December 2012), pp. 1165-1188 (24 pages) https://doi.org/10.2307/41703503.

Kiefer, C. (2019). "Quality Indicators for Text Data", H. Meyer et al. (Hrsg.): BTW 2019— Workshopband, Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019, 145.

Esteves, D., Rula, A., Reddy, A. J. & Lehmann, J. (2018). "Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis", J. Data and Information Quality 9, 3, Article 16 (February 2018), doi: 10.1145/3177873

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Managemen*t, 35(2), 137–144.

Gupta, S. & Chaudhari, M. S. (2015). Big Data Issues and Challenges: Data analysis, storing, processing, issues, challenges and future scope‖, International Journal on Recent and Innovation Trends in Computing and Communication, vol. 3, Issue 2, pp. 62-67, February 2015.

Gwenaelle C. S. & Lee, M. (2021). "Stacked DeBERT: All Attention in Incomplete Data for Text Classification", Neural Networks, January 15, 2021.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T. & Yenanegi, M. R. (2020). Big Data and Cognitive Computing.

Howraa M. M. A., Kod, M. S., Al-Salim, A.M.A., & Al-amiri, C. (2020). "Algorithm for Greening Big Data Networks in Iraq Under Veracity Dimension", 3rd International Conference on Engineering Sciences, IOP Conf. Series: Materials Science and Engineering 671 (2020) 012052, IOP Publishing, doi:10.1088/1757-899X/671/1/012052.

Espinosa, J. A., Kaisler, S., Armour, F., & Money, W. H. (2019). "Big Data Redux: New Issues and Challenges Moving Forward", Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

Keskar, V., Yadav, J., & Kumar, A. H. (2020). 5V's of Big Data Attributes and their Relevance and Importance across Domains, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-9 Issue-11, September 2020.

Kotani, K., Yoshimi, T., & Hitoshi I. (2011). "A Machine Learning Approach to Measurement of Text Readability for EFL Learners Using Various Linguistic Features", US-China Education Review B 6 (2011) 767-777, ISSN 1548-6613.

Subramaniam, L. V., Roy, S., Tanveer A. & Faruquie, Negi, S. (2009). 'A Survey of Types of Text Noise and Techniques to Handle Noisy Text', DOI: 10.1145/1568296.1568315 · Source: DBLP.

Lacasta, J. Nogueras-Iso, J. Francisco, J. & Soria, Z. (2010). Terminological Ontologies: Design, Management and Practical Applications, Springer Science & Business Media, 3 Aug. 2010 - Computers - 198 pages.

Lozano, M. G., Brynielssona, J., Frankec, U., Rosella, M., Tjörnhammar, E., Varga, S., & Vlassov, V. (2020). "Veracity assessment of online data", Decision Support Systems 129 (2020) 113132, Science Direct, Elsevier.

Martinez-Gil, J. (2023). Optimizing Readability Using Genetic Algorithms https://arxiv.org/pdf/2301.00374.pdf, https://doi.org/10.48550/arXiv.2301.00374.

Mikalef, P., Pappas, & I.O., Krogstie, J. *et al.* (2018). Big data analytics capabilities: a systematic literature review and research agenda. *Inf Syst E-Bus Manage* 16, 547–578 (2018). https://doi.org/10.1007/s10257-017-0362-y.

Kiyavitskaya, N., Zeni, N., Mich, L. *et al.* (2008). Requirements for tools for ambiguity identification and measurement in natural language requirements specifications. *Requirements Eng* **13**, 207–239 (2008).

https://doi.org/10.1007/s00766-008-0063-7
Nimmagadda, S.L., Zhu, D. and Reiners, T. (2019). On Managing Contextual Knowledge of Digital Document Ecosystems, characterized by Alphanumeric Textual Data, *Procedia Computer Science*, Volume 159, 2019, pages 1135-114, https://doi.org/10.1016/j.procs.2019.09.282.
Omidbakhsh, M. & Ormandjieva, O. (2020). Toward a New Quality Measurement Model for Big Data. DOI: 10.5220/0009820201930199 In Proceedings of the 9th International Conference on Data Science, Technology and Applications (DATA 2020), pages 193-199.
Oseguera, O., Rinaldi, A., Tuazon, J., & Cruz, A.C. (2017). Automatic Quantification of the Veracity of Suicidal Ideation in Counseling Transcripts. In: Stephanidis, C. (eds) HCI International 2017 – Posters' Extended Abstracts. HCI 2017. Communications in Computer and Information Science, vol 713. Springer, Cham. https://doi.org/10.1007/978-3-319-58750-9_66.
Ylijoki, O., & Porras, J. (2016). "Perspectives to Definition of Big Data: A Mapping Study and Discussion", *Journal of Innovation Management,* Ylijoki, Porras JIM 4, 1 (2016) 69-91.
Padillo, F., Luna, J.M. & Ventura, S. A. (2019). Grammar-Guided Genetic Programing Algorithm for Associative Classification in Big Data. *Cogn Comput* 11, 331–346 (2019). https://doi.org/10.1007/s12559-018-9617-2.
Rainer, K. & Prince, B. (2022). Introduction to Information Systems, supporting and transforming business, ninth edition, John Wiley & Sons, International Adaptation, New Jersey, USA.
Reihaneh H. H., Fredericks, E. M. & Bowers, K. M. (2019). "Uncertainty in big data analytics: survey, opportunities and challenges", Hariri et al. J Big Data (2019) 6:44, doi:10.1186/s40537-019-0206-3.
Crone, R. (2016). Big Data Veracity Assessment, improving risk assessment by adding high veracity data to existing contents insurance models, Delft University of Technology, 2016.
Rudra, A. & Nimmagadda, S.L. (2005). Roles of multidimensionality and granularity in data mining of warehoused Australian resources data, *Proceedings of the 38th Hawaii International Conference on Information System Sciences*, Hawaii, USA.
Gupta, S., & Gupta, A. (2019). "Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review", *The Fifth Information Systems International Conference* 2019, Science Direct, Procedia Computer Science 161 (2019) 466–474.
Srivastava, A., & Sahami. M. (2009). *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL: CRC Press. ISBN 978-1-4200-5940-3.
Sivarajah, U., Irani, Z., & Weerakkody, V. (2015). Evaluating the use and impact of Web 2.0 technologies in local government. Government Information Quarterly, 32(4), 473–487.
Sivarajah, U., Kamal, M.M., Irani, Z. & Weerakkody, V. (2017). 'Critical analysis of Big Data challenges and analytical methods', Journal of Business Research 70 (2017) 263–286.
Sowmya, V. (2021). Trends, Limitations and Open Challenges in Automatic Readability Assessment Research.
Vasarhelyi, M. A., Kogan, A., & Tuttle, B. M. (2015). Big data in accounting: an overview. *Accounting Horizons*, 29(2), 381–396.
Wu, X., Zhu, X., Wu, G. Q. & Ding, W. (2014). "Data mining with big data," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97-107, Jan. 2014, doi: 10.1109/TKDE.
Wang, X., Luoa, X., & Liu, H. (2014). "Measuring the veracity of web event via uncertainty", *The Journal of Systems and Software*, G Model JSS-9355; No. of Pages 11, Science Direct, Elsevier.
Liu, Y., Ji, M., Lin, S. S., Zhao, M. & Lyv, Z. (2021). "Combining Readability Formulas and Machine Learning for Reader-oriented Evaluation of Online Health Resources", IEEE Access, DOI:10.1109/ACCESS.2021.3077073.
Yang, Y. (2019). "Uncertainty and grey data analytics", *Marine Economics and Management*, Vol. 2 No. 2, 2019, pp. 73-86, Emerald Publishing Limited, 2516-158X, DOI 10.1108/MAEM-08-2019-0006.
Zhang, Y., Lin, N., & Jiang, S. (2019). "A Study on Syntactic Complexity and Text Readability of ASEAN English News", IALP 2019, Shanghai, Nov 15-17, 2019.