Semantic Text Mining using Domain Ontology

Ibukun T. Afolabi, Olaperi Y. Sowunmi and Taiwo Adigun

***Abstract*— Presently in Customer Relationship Management, there is a need to achieve greater customer centricity, and this requires a deeper understanding of customer needs. Also, the volume of textual data generated by the social networking sites in recent times has greatly increased, creating a platform for analysis, towards the much needed customer understanding. One of the issues that evolve from analyzing these texts to retrieve non trivial patterns (text mining) is text representation, which this research is aimed at addressing. In particular, this paper focuses on using domain ontology for text pre-processing in order to improve the quality of the textual corpus being mined. The methodology used in this research is based on developing a domain Ontology for textual pre-processing of the experimental data and sentiment analysis of social media data. In conclusion, the inferences gotten from the research carried out reveal that domain ontology has the ability to improve the results of sentiment analysis. It was also discovered that, due to the nature of social media data, there is need for a deeper level of semantic analysis, to be able to maximize its richness.**

***Index Terms*— Domain Ontology, Text Mining, Banking, Social Media, Sentiment Analysis, Slangs, acronyms**

I. Introduction

According to [1], social networking can be described as convergence of technologies that makes it possible for individuals to easily communicate, share information, and form new communities online. Some popular examples of social networking sites include but not limited to Facebook and Twitter [2]. Mark Zukerberg and his roommate founded Facebook in 2004 [3]. Twitter on the other hand, is a status update that allows people to share short updates about people or events and to view updates created by others [4]. Twitter is a fast growing micro blogging service with over 320 million users as of January 2016. Twitter users tweet about any topic within the 140-character limit and follow others to receive their tweets [5]. Recently, social networks have experienced huge amount of growth in the number of users [6], this has created a corresponding increase in amount of unstructured or textual data available for analysis. This is principally because the mode of communication on these social networks is primarily through chatting, which involves typing textual comments. Along with this recent development is the fact that, social networks have changed the way in which customers relate with companies or business organizations generally. In particular, customers have gained more and more control over and through the communication regarding the company and its products [1]. One important opportunity that social textual data can provide in customer relationship management is customer service, for example, social media text can be tagged, and interesting commentary e.g. “possible recall,” “product defect,” ”confusing instructions”, can be discovered to improve marketing. Also capturing social media conversations in context, provide more robust information. In order to maximize the opportunity described above, it is important to introduce semantic analysis into the process of text analysis. According to [7], some of the issues that evolve from analysing text to retrieve non trivial patterns(text mining) include dealing with noisy data, word sense disambiguation, text tagging, text representation i.e which terms are important?, what about word order, context and background knowledge? Furthermore, [7] concluded that taking into account the language a text is written is very important, since the language highlights the morphological analysis need. Also the domain of a text collection determines the difference between the technical terms and redundant terms. According to [8], one of the main issues in research automatically understands noisy social media text. This is because it may contain a mixture of more than one language, slangs, and acronyms and so on. In addition, even though social media text has huge benefits, it also introduces some challenges for information access and language technology [8]. Some of these challenges include the fact that, social media text is characterized by having a high percentage of spelling errors and containing creative spellings (“gr8” for ‘great’), phonetic typing, word play (“goooood” for ‘good’), abbreviations (“OMG” for ‘Oh my God!’), Meta tags (URLs, Hashtags), and so on [8].

Most of the text mining systems reviewed have concentrated on English texts especially in the area of text pre-processing, however in this paper, we propose a text mining text pre-processing stage which uses slang and acronyms ontology. The idea is to investigate if there is an improvement in the mining process using the proposed approach.

Manuscript received July 1, 2017; revised July 23, 2017. This work was supported by Covenant University. I.T. Afolabi, O.Y. Sowunmi and Taiwo Adigun. thanks Covenant University for their Sponsor and financial support towards the success of this publication.

I.T. Afolabi is a faculty in the department of Computer and Information sciences, Covenant University, Ota, Nigeria(corresponding author to provide phone: +234- 08021247616; e-mail:ibukun.fatudimu@covenantuniverity.edu.ng).

O. Y. Sowunmi is a faculty in the department of Computer and Information sciences, Covenant University, Ota, Nigeria (e-mail: olaperi.sowunmi@covenantuniversity.edu.ng)

Taiwo Adigun is a faculty in the department of Computer and Information sciences, Covenant University, Ota, Nigeria (e-mail: taiwo.adigun@covenantuniversity.edu.ng)

II. Literature Review

Text mining is a multidisciplinary field comprising “ information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning and data mining”, it can be

viewed as an extension of data mining and knowledge discovery in databases [9]. It is the process of analysing and discovering meaningful and useful information from natural language text which is usually semi-structured or unstructured and relatively difficult to handle algorithmically [10]. However, the potential value of the output makes the process worthwhile. Text mining is concerned about searching out and extracting patterns in unstructured text by identifying and exploring interesting patterns [11,12]; as a subset of text analytics, it focuses on applying Natural Language Processing (NLP) and machine learning to textual information. Its analysis is based only on the syntax of the given text, that is, the structural relationship between the words and not on the pragmatics, discourse or phonetics [13].

Text mining has been successfully applied in different areas of endeavours including the business environment, security – for fraud detection and crime prevention, in hospitals and pharmaceutical companies and in the telecommunication industry [13].

A lot of mining has been carried out on social media data, moreover, it is greatly being used by companies to understand the conception of their customers and by so doing, develop and improve their competitive strategy and thus boost financial performance [14].

Ref. [15] worked on the detection of events and event content identification from social media data and proposed techniques to handle noise and heterogeneity of such data. The authors applied the techniques on sample data from Flickr, and it proved effective. Ref. [16], on the other hand investigated and improved on text-based geo-location prediction of users based on their tweets. Ref. [14] used a text mining tool from *socialmention.com* to analyse text from primary social media sites for about 83 hotels to understand sentiments, passion and reach. The study investigated the effect competitive intelligence gained via text mining has on the improvement of Return on Equity (ROE) of companies, and found it to be positive. Ref. [17] extracted attribute and value pairs from product descriptions that were captured in text format. They applied single-view and multi-view semi-supervised learning algorithms to extract both implicit and explicit attributes in the product description. Naïve Bayes and the Expectation-Maximization algorithms were used for labelled and unlabelled data respectively. They also showed how resulting information can be used as information for recommender systems and competitive intelligence tools.

Traditional data mining usually requires some domain knowledge to get the best results, however this is usually captured manually by human domain experts. Semantic data mining thus augments the traditional data mining process by deriving domain knowledge from semantic data and incorporating it into the process; this domain knowledge serves as background knowledge which are very useful. Semantic web technologies such as the Resource Description Framework (RDF) and Web Ontology Language (OWL) enable this knowledge to be captured automatically with minimal effort, rather than entirely manually by domain experts [18]. Ref. [19, 20, 21, 22] applied semantics mining to text mining via the use of ontologies and [19, 21, 22] applied same specifically to improve rules generated by association rule mining.

Ref. [20] presented an integrated and ontology-enriched framework for link prediction in social networks. The authors upgraded a previous work on the analysis of friendship networks which was limited by the flat representation used with ontology enriched features. Tests on the integrated approach yielded better results on the precision and recall of known friendship. Ref. [18] also proposed integrating semantic data mining via the use of ontologies into an already existing data mining system, showing how the performance of classification and knowledge extraction would improve.

With the aim of reducing the 'search space' of the algorithm and improve the significance of the rules generated, [19] proposed an integrated framework to extract "constraint-based multi-level association rules with the support of an ontology. The framework was tested with a market basket analysis application. Ref. [21] on the other hand researched on a technique to improve the post-processing stage of association rule mining via the use of ontology to reduce the large number of rules extracted by the Apriori algorithm. The authors integrated domain knowledge and also generalized general impressions with the use of Rule Schemas. Ref. [22] carried out a similar research to [21] and experimental results showed the efficiency of the new approach in reducing the number of rules. The authors proposed the use of schemas and domain ontologies to prune and filter rules discovered from association rule mining. The domain ontology was used to integrate and strengthen user knowledge in the post-processing phase

III. Methodology

3 Methodology

The methodology used in this research is based on the following major steps:

i. Gathering of Textual data used for the experiment.
ii. Developing a domain Ontology for Textual pre-processing of the experimental data
iii. Pre-processing of Textual data used for the experiment.
iv. Knowledge distillation using Sentiment analysis.
v. Comparing the result generated by the ontology based text mining and non-ontology based text mining.

The activity work flow for the research methodology is presented in Figure 2.

3.1 Data Gathering

Data was gathered for the experiment from twitter (*https://twitter.com/gtbank)* account of the bank used as case study which is Guaranty Trust Bank (GTB). The data was gathered within the period of July 1st to 31st, 2015. A total of 5,934 tweets were gathered. According to [23], social media is a top trend in the banking industry. It was also revealed that social media has an inevitable role in the banking industry to retain the brand loyalty of particularly digitally savvy customers. Also social media is increasingly becoming a popular media outlet among consumers with over 89% of customers surveyed stating they had a social media account. This makes it necessary for the banking

ISBN: 978-988-14048-7-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

industry to intensify the integration of social media trends with banking services and also analyse social media content for competitive advantage. GTBank (Nigeria) launched its innovative 'Social Banking' service on Facebook, being the first in Nigeria, with over 950,000 Facebook fans. Although, the social account differs from a regular GTBank account, the new channel allows GTBank social account holders to transfer money, purchase airtime, pay bills, and confirm their account balance on Facebook. As the social media trend grow, more banks are expected strategically setup and maintain social media presence to engage their customers, as some banks in Nigeria (e.g. GTBank) have utilized social media in advancing what is termed 'social banking'; with a considerable followership base.

3.2 Developing the Domain Ontology

The domain ontology developed for the experiment is a slang ontology. It contains slangs and acronyms commonly used on the social media. In addition to this, it also contains pidgin English which is common to the textual contents of Nigerian social media page. In pidgin English, some of the English words have been changed, for example "is" is being referred to as "na", "will not" is being referred to as "no go" etc. The core idea of the approach to developing the slang ontology is based on the prescriptions in [24, 25]. Relevant keywords and key-phrases were extracted from textual information using the typical information extraction process [25]. The ontology was built using the following steps: 1) Extract important terms from text: 2) define the concept taxonomies, relations, attributes, instances, axioms and functions: To implement the slang ontology, the protégé ontology editor was used. The slang ontology contains the following constructs;

rdfs:label: Used to store any other known reference to the terms addressed by the ontolgoy. **rdfs:** Comment: Captures the further explanation on the term or individual. ***owl:sameIndividualas:*** used to equate individuals with similar meaning for example "cuz" is equated to "because", ***rdfs:subClassOf*** **:** Used to break down general concepts to the type of classes that make them up. The slang ontology contains relevant Object properties, such as *HasSynonymn, HasAsExample, PartOf,* all of which define relationships between concepts in the ontology. The terms in the ontology is gotten from various sources, such as websites, pidgin English translators etc. The slang ontology contains 5 concepts, 3 object properties, 898 individuals. The concepts include "bankslang", which is a collection of terms pertaining to the banking industry, "EnglishAcronyms" is a collection of general English acronyms, "Interpretation" contains the English language interpretation of the slangs, acronyms etc, "negativeSlang" contain slangs and acronyms used to depict negative words, abusive works, vulgar words etc, "positive Slang" contains appreciation, encouragement etc while every other slang and acronymns fall under the "neutralSlangs". Figure 1 is a snapshop of the slang ontology. An application that replaces slang (using the slang ontology) with the English language interpretation is then developed using the Java programming language.

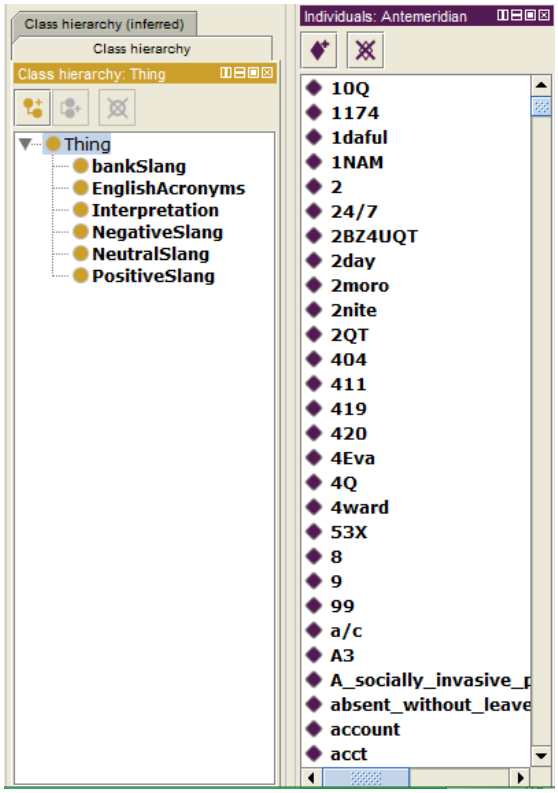

Figure 1: Screen shot of the slang and acronyms ontology

3.3 Pre-processing Textual Data

The following step was carried out to pre-process the data

1. Filtration: This process has to do with removing the unimportant words from documents content. Such unimportant words include: articles, pronouns, determiners, prepositions and conjunctions, common adverbs and non-informative verbs. As a result of this process, more important or highly relevant words are singled out.
2. Stemming: This is the process that removes a word's prefixes and suffixes (such as unifying both infection and infections to infection). For this experiment, the stemming algorithm referenced the wordNet Dictionary (https://wordnet.princeton.edu/wordnet/download/current-version/) in order to order to have a level of semantic analysis introduced into this pre-processing.
3. Finally, the weighting scheme TF-IDF (Term Frequency, Inverse Document Frequency) is used to assign higher weights to syntactically distinguished terms in a document [26]. All the above was carried out using the rapidminer toolkit [27].

3.4 Sentiment Analysis

"What other people think" has always been an important piece of information for most of us during the decision-making process. Sentiment analysis, also called opinion mining, analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It focuses on opinions which express or imply positive or negative sentiments [28]. This research uses the "Extract Sentiment" operator in rapid miner studio to extracts sentiment based on SentiWordNet 3.0.0, from the individual tweets for GTBank. This operator uses a WordNet 3.0 and a SentiWordNet 3.0.0 database to

ISBN: 978-988-14048-7-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

extract sentiment of an input document. The sentiment value is in range [-1.0,1.0] where -1.0 means very negative and 1.0 means very positive. WordNet and SentiWordNet are connected by Synset IDs. To determine the sentiment of a document we compute sentiment of each word, where the first meaning of a word has the most influence on a sentiment and each next meaning has less influence on a sentiment. Document sentiment is then computed as the average value of all word sentiments [27].

3.5 Comparing Result

After applying sentiment analysis on the data pre-processing with and without the slang ontology, the results were compared to discover the effect of slang ontology on the quality of the results gotten.

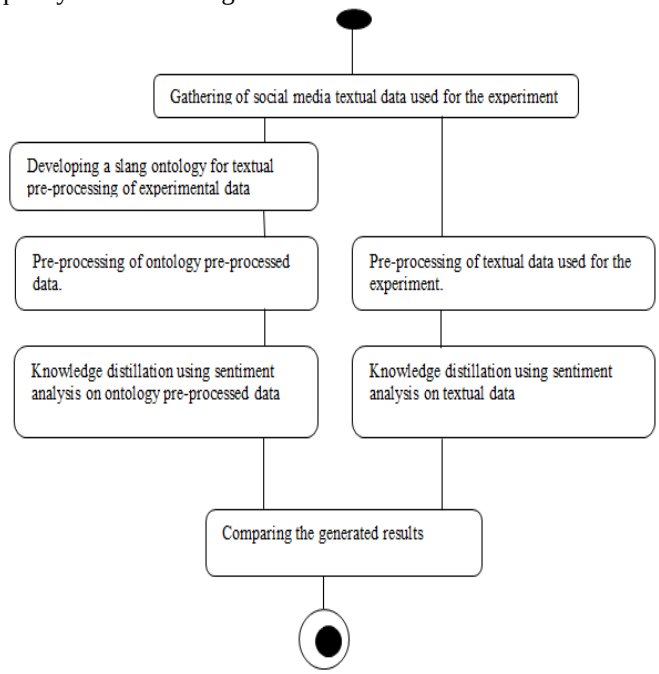

Figure 2: Workflow diagram of the methodology

IV. RESULT AND DISCUSSION

The total tweets for GTBank in the month of July 2015 were combined and sentiment analysis was performed on the resulting textual corpus to generate sentiment values results for each tweet. The results are visualized in the graph shown in Figure 3. Figure 3 is the combination of analysis of the tweets pre-processed using the slang ontology and the tweets that were not pre-processed using the slang ontology. The figure reveals a significant difference in the sentiment value of some tweets when they were pre-processed. Some tweets showed significant increase in negative value after the ontology pre-processing while some also increase in positive value, though most of the tweets still retained their value.

Figure 4 contains the most popular slangs that were replaced in the ontology pre-processed file. The some of the replaced terms in the pre-processed textual document are stopwords for example “your” is “ur”, “you” is “u” etc there are also important keywords such as plz-please , atm-at the moment, lol-laugh out loud, which contribute to the sentiment analysis.

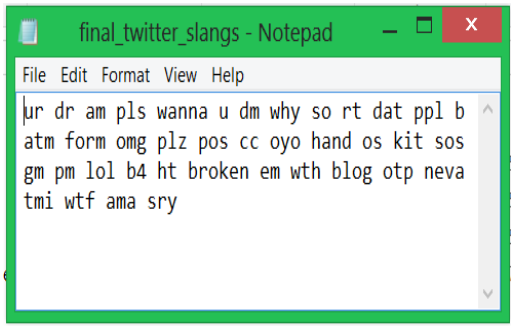

Figure 4: Extracted Slang and acronyms

Also, on a percentage basis, the percentage of tweets that their sentiment value was either altered positively or negatively was 43%. This reveals the fact these slangs, acronyms, pidgin English etc used on the social media do have effect on the sentiment analysis result, especially when applied in the social media context.

Also discovered generally as regards applying sentiment analysis on the GTBank tweets in the month of July in 2016 is that, for the ontology based pre-processed tweets the most positive sentiment value is 0.784 while the non-ontology pre- processed tweets had 0.875 as the most positive sentiment value. Also, the ontology based pre-processed tweets had the most negative sentiment value is -0.75, while the non-ontology pre- processed tweets also had -0.75 as the most negative sentiment value.

V. CONCLUSION AND FUTURE WORK

IIIn Conclusion, this paper has investigated a text mining approach for sentiment analysis using a slang domain ontology. The research was able to compute the contributions introduced into the mining process by the slang domain ontology. Also, the inferences gotten from the research carried out real that domain ontology has the ability to improve the results of sentiment analysis. In addition to have a complete interpretation and analysis of social media data, there is a level of semantic analysis required. The developed ontology will also be a useful resource in investigating the Nigerian social media content for detecting interesting issues such as security treats, abnormal behaviour of teens online and so on. This is particularly interesting in that these kinds of contents contain so much more of these slangs than the banking social media site investigated in this research. In order to improve this research, further work can be done to, add more contents to the ontology and make it more robust. By increasing the domain covered by the slang ontology, such research will be able to address a wide range of issues. Also, other mining algorithms like association rule and k-means etc can be investigated and also other social media sites such as facebook etc.

ISBN: 978-988-14048-7-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

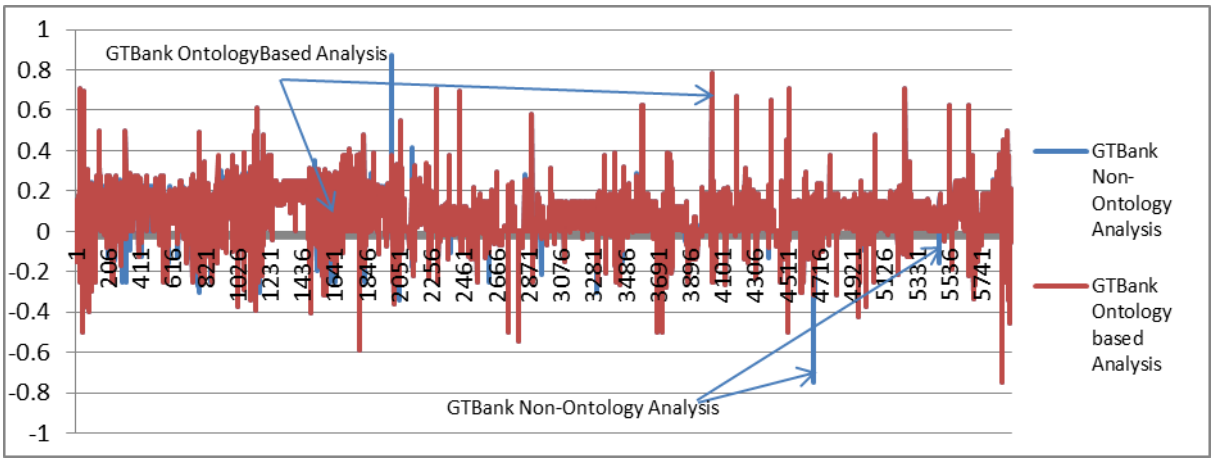

Figure 3: Sentiment Analysis Result

REFERENCES

[1] W. Assaad. and J.M. Gómez "Social Network in marketing (Social Media Marketing) Opportunities and Risks" in International Journal of Managing Public Sector Information and Communication Technologies (IJMPICT) Vol. 2, No. 1., 2012.

[2] M. Dewing. 'Social Media: An Introduction'. *Library Parliament Publication* No. 2010-03-E, 2012

[3] F. Farooq, Z. Jan and S. Karachi 'The Impact of Social Networking to Influence Marketing through Product Reviews'. *International Journal of Information and Communication Technology Research, 2012*

[4] S. Fox, K. Zickuhr and A. Smith (2009). 'Twitter and Status updating'. *Pew internet and American life project. (http://www.pewinternet.org/2009/02/12/twitter-and-status-updating/)*

[5] H. Kwak, C. Lee, H. Park and S. Moon, (2010) 'What is Twitter, a social network or a news media?', *Proceedings of the 19th international conference on World wide web, April 26-30, 2010, Raleigh, North Carolina, USA* [doi>10.1145/1772690.1772751]

[6] S. Stieglitz, L. Dang-Xuan, A. Bruns and C. Neuberger "Social Media Analytics. An Interdisciplinary Approach and Its Implications for Information Systems" in Business & Information Systems Engineering. DOI 10.1007/s12599-014-0315-7, 2014

[7] A. Stavrianou, P. Andritsos and N. Nicoloyannis, "Overview and Semantic Issues of Text Mining" in SIGMOD Record, Vol. 36, No. 3, Pp23 -34, 2007

[8] A. Das and B.Gamback. "Code-Mixing in Social Media Text: The Last Language Identi_cation Frontier?" Traitement Automatique des Langues (TAL): Special Issue on Social Networks and NLP , TAL Volume 54 no 3/2013, Pages 41-64, 2014

[9] A.H.Tan. "Text mining: The state of the art and the challenges". In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases* (pp. 65–70), 1999

[10 I. H. Witten. Text Mining. In M. P. Singh (Ed.), "*The Practical Handbook of Internet Computing"* (pp. 14–1 – 14–22). Boca Raton, Florida: Chapman and Hall/CRC. http://doi.org/10.1201/9780203507223.ch14, 2005

[11] R. J. Mooney and Nahm U. Y. "Text Mining with Information Extraction". In W. Daelemans, T. du Plessis, C. Snyman, & L. Teck (Eds.), *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium (Studies in Language Policy in South Africa 4)* (pp. 141–160). Bloemfontein, South Africa: Van Schaik Publishers, 2005

[12] M. Radovanovic and M. Ivanovic, "Text Mining: Approaches and Applications" *Novi Sad J. Math.*, *38*(3), 227–234, 2008

[13] G. Chakraborty, M. Pagolu and S. Garla. "*Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS"*. SAS Institute, 2013

[14] G. Gémar and J.A. Jiménez-Quintero. "Text mining social media for competitive analysis". *Text Mining Social Media for Competitive Analysis*, *11*(1), 84–90, 2015

[15] F. Psallidas, H. Becker, M. Naaman & L. Gravano. "Effective Event Identification in Social Media". *IEEE Data Eng. Bull*, *36*(3), 42–50, 2013

[16] B. Han, P. Cook & T. Baldwin. "Text-Based Twitter User Geolocation Prediction". *Journal of Artificial Intelligence Research*, *49*, 451–500, 2014

[17] R. Ghani, K. Probst, Liu, Y., Krema, M. and Fano, "A. Text Mining for Product Attribute Extraction". *ACM SIGKDD Explorations Newsletter*, *8*(1), 41–48, 2006

[18] F. Benites, F., & E. Sapozhnikova "Using Semantic Data Mining for Classification Improvement and Knowledge Extraction". *LWA*, 150–155, 2014

[19] A. Bellandi, B. Furletti, V. Grossi and A. Romei, "Ontology-Driven Association Rule Extraction: A Case Study" In *Contexts and Ontologies Representation and Reasoning* (p. 10). Roskilde, Denmark, 2007

[20] W. Aljandal, V. Bahirwani, D. Caragea and W. Hsu, "Ontology-Aware Classification and Association Rule Mining for Interest and Link Prediction in Social Networks". *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0.*, 3–8, 2009

[21] C. Marinica and F. Guillet "Improving Post-Mining of Association Rules with Ontologies". In L. Sakalauskas, C. Skiadas, & E. K. Zavadskas (Eds.), *The XIII International Conference "Applied Stochastic Models and Data Analysis" (ASMDA-2009)* (pp. 76–80), 2009

[22] P.S. Bhavani, M.N.H. Bindu and K.S. SundaraRao "Ontology Based Post Mining of Association Rules". *International Journal of Innovative Research in Science, Engineering and Technology*, *4*(7), 5179 – 5188, 2015

[23] V. K. Suvarna and B. Banerjee. "Social Banking: Leveraging Social media to enhance customer engagement. Capgemini", 2014

[24] S. Boyce and C. Pah, "Developing Domain Ontologies for Course Content". In *Educational Technology & Society*, 10 (3), 275-288, 2007

[25] N.F. Noy and D.L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology". http://protege.stanford.edu/publications/ontology_development/ontology101.pdf , 2000

[26] M. Hany, R. Dietmar, I. Nabil. and T. Fawzy. "A Text Mining Technique Using Association Rules Extraction", International Journal of Computational Intelligence volume 4 number 1 2007 ISSN 1304-2386, 2007.

[27] RapidMiner Studio Manual (Accessed 2014) (https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf).

[28] L. Bing 'Sentiment Analysis and Opinion Mining'. *Synthesis Lectures on Human Language Technologies.* Morgan & Claypool Publishers, 2012

ISBN: 978-988-14048-7-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)