A Government Decision Analytics Framework Based on Citizen Opinion (Gov-DAF): Elaboration of the Knowledge Base Component

Mohamed Adel Rezk[1][2], Adegboyga Ojo[2], Ghada El Khayat[1], Safaa Hussien[1]

[1]Department of Information Systems and Computers,
Alexandria University,
Alexandria, Egypt
ghadaek@gmail.com, safaa26bedo@yahoo.com

[2]Insight Centre for Data Analytics,
National University of Ireland,
Galway, Ireland
mohamed.rezk@nuigalway.ie, adegboyega.ojo@nuigalway.ie

ABSTRACT

Citizens' satisfaction index towards public policies is a core political research question. Seeking agile and efficient public policies, public policy makers are perpetually investigating how to measure citizens' satisfaction towards their policies, in order to overhaul faulty public policy aspects or topics that outcome a negative citizens' satisfaction index. Bearing in mind that a correct calculation will grant a significant success for the public policy and the public policy makers. Our perspective is that nevertheless the index was well calculated or not it is always too late, public policy is already issued and citizen reactions are ensue. Hence, we previously proposed a public policy satisfaction prediction framework. This framework reckons on a knowledge base that allows formulating the prediction formulae. To develop this knowledge base we extend the Core Public Policy Vocabulary (CPPV), apply Named Entity Recognition and Topic Modeling, in parallel, for keywords extraction and semantic similarity measuring in order to relate the detected keywords with the pre-defined public policy aspects. The aforementioned methods will allow an automated population of the prediction knowledge base.

Keywords— Citizens' Satisfaction Index; Core Public Policy Vocabulary; Public Policies; Policy Satisfaction Prediction; Citizen Reactions; Topic Detection; Semantic Similarity; Automated Ontology Population; Policy Aspects.

I. INTRODUCTION

To achieve citizen's satisfaction measurement multiple automated and manual citizen's' satisfaction index calculation methods are applied [1]–[10], In 2012 Bandari et al [11] introduced their news article popularity forecasting model. In this model, the news article popularity variable is measured by the number of news article url sharing in twitter. They used four independent variables to build the predictive model; those variables are news source, news category, article subjectivity and named entities mentioned in the news article. In our Government Decision Analytics Framework Based on Citizen Opinion (Gov-DAF), we are building two predictive models for forecasting the actual public policy acceptance rate as a dependent variable. This variable is quantified using Actual Satisfaction Rate function *Fig. 1*, against the independent variable in this case which is Social Media Public Policy Acceptance quantified using Micropost Satisfaction Rate function that uses the sentiment analysis of the tweets as the scoring method *Fig. 1*.

Gov-DAF is contributing and extending multiple algorithms and tools for building a solution to analyze the tweets sentiments in order to solve the Micropost Satisfaction Rate equation *Table 1*.

$$\text{Microposts satisfaction rate} = \frac{\sum_{k=0}^{k=n} \text{positive tweets}}{\sum_{k=0}^{k=n} \text{positive tweets} + \sum_{k=0}^{k=n} \text{negative tweets} + \sum_{k=0}^{k=n} \text{neutral tweets}}$$

$$\text{Actual satisfaction rate} = \frac{\sum_{k=0}^{k=n} \text{positive survey}}{\sum_{k=0}^{k=n} \text{positive surveys} + \sum_{k=0}^{k=n} \text{negative surveys} + \sum_{k=0}^{k=n} \text{neutral surveys}}$$

Fig. 1. Mathematical Representation of the Gov-DAF Analysis Model
(k = keyword within public Policies)

TABLE I. GOV-DAF [12], [13]

Public Policy Ontology Modeling	Computational Analysis of Citizen Opinions and Sentiments	Knowledge base Population	Prediction
• Invistigated public policy. • Origin and branch keyword extraction from Policy document. • Linking keywords to policy aspect.	• Harvesting citizen contents from Twitter. • Opinion mining.	• Populating our ontology with keywords. • Attaching Citizen Opinions.	• Mining the accumulated knowledge to calculate citizen satisfaction rates towards policy aspects. • Producing citizen satisfaction insights.

Gov-DAF knowledge base pipeline implementation adopted the following vocabulary and two methodologies: (1) CPPV [14]. (2) Initial approach “Named Entity Recognition Based Methodology” of building the Gov-DAF knowledge base pipeline as introduced in [13], [12]. (3) Enhanced approach “Topic Modeling Based Methodology”. Finally we evaluated both methodologies against each other and results are presented in this paper.

Through both methodologies the following vocabularies, algorithms and tools are extended and/or adopted as detailed below:

A. *Structured Public Policy Indexing using CPPV and CKAN*

Gov-DAF reckons on the mined keywords exists in public policies as assets for achieving the ultimate target of obtaining meaningful insights about public policies that help public policymakers. Thus, semantically structured collection and indexing of public policies or assets for Gov-DAF analytical purposes was one of the main motivations of creating CPPV [14] as part of the research. CPPV offers semantic indexing of public policies’ (assets) metadata *Fig. 2*, and CKAN “Comprehensive Knowledge Archive Network” [15] the world’s leading open-source data portal platform offers public policy physical documents indexing.

Within Gov-DAF Knowledge base building pipeline we extended CPPV with the public policy analytics classes (cppv-ext:AnalyticalAspect, :Keyword), and properties (cppv-ext:type, :occorance_count, :extends, :composed_of) as the Gov-DAF knowledge base elements that will enable opinion harvesting and analysis in later phases *Table 1*.

B. *Public Policy Text Analysis by Named Entity Recognition using Stanford NER and Topic Modeling using LDA*

To populate Gov-DAF knowledge base with keywords extracted from public policies we applied two text analysis methods and measured their accuracy indices in our usage domain, which does not necessarily indicate their overall accuracy in other usage domains. This is discussed in section 3.

The first method applied for public policy text analysis is Named Entity Recognition using stanford NER [16] where Gov-DAF uses stanford NER to extract persons, places and organizations that are composing the public policy. The second method applied for public policy text analysis is Topic Modeling using Mallet implementation of LDA “Latent Dirichlet Allocation” [17]. Here Gov-DAF applies Mallet LDA to cluster keywords composing the public policy into topics vector and then NER is applied to discover keywords types with the possibility to apply Stanford Entity Resolution Framework [18].

C. *Semantic Relatedness using DISCO*

Semantic relatedness is used in two cases in this work. The first use case is when NER Methodology is applied as Gov-DAF uses Semantic Relatedness through Extracting DIStributionally related words using CO-occurrences (DISCO) [19], [20] for generating branch keywords *Fig. 4*.

The second use case is just before applying opinion harvesting, sentiment analysis and satisfaction estimation according to Gov-DAF *Table 1*. Gov-DAF first relates keywords back to certain public policy aspects for deeper

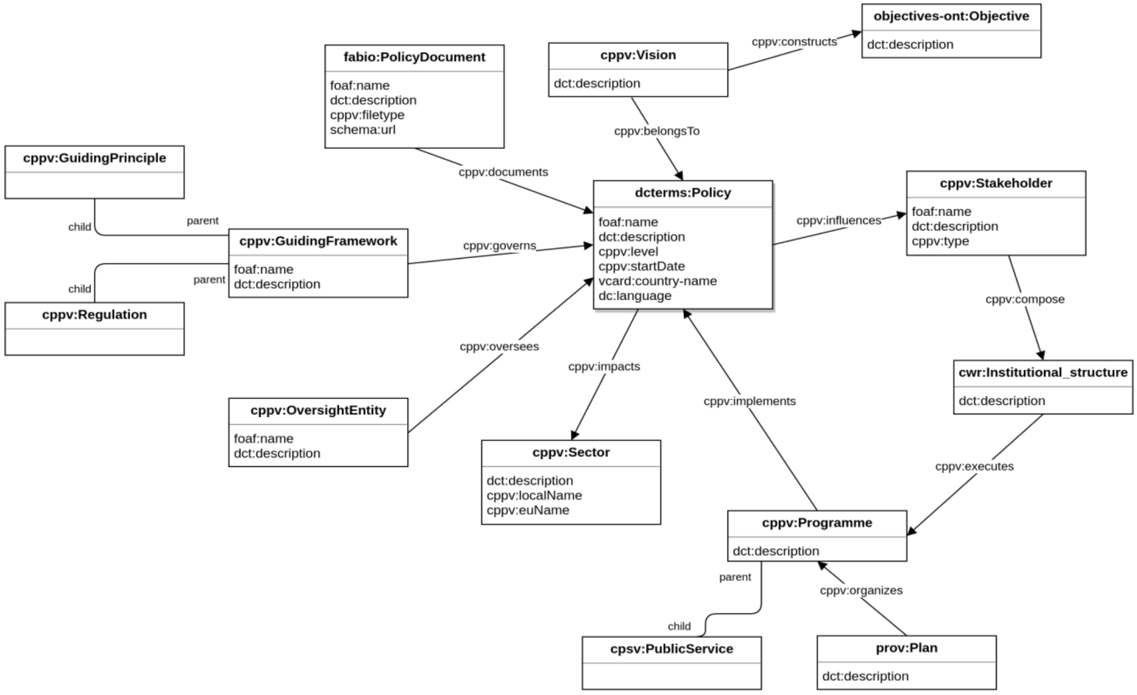


Fig. 2. CPPV Conceptual Design [14]

public policy analysis. In other words, for a multi-level satisfaction estimation, Gov-DAF applies Semantic Relatedness using DISCO "Extracting DIstributionally related words using CO-occurrences" for the target of relating extracted public policy keywords with public policy aspects using the semantic relatedness based algorithm illustrated in [13], [12].

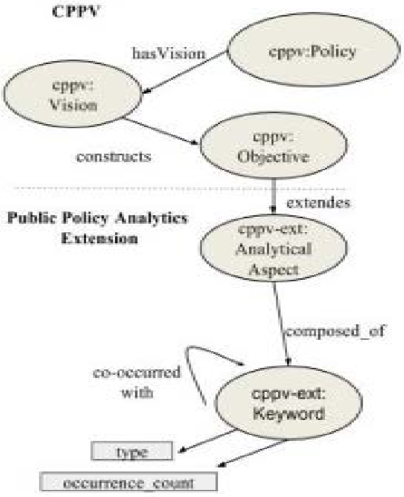


Fig. 3. CPPV Policy Analytics Extension

II. GOV-DAF KNOWLEDGE BASE BUILDING PIPELINE

Gov-DAF follows the initial approach "Named Entity Recognition Based Methodology" [13], [12], and also the new enhanced approach "Topic Modeling Based Methodology" for designing Gov-DAF knowledge base building pipeline. Named Entity Recognition based Methodology was the initial methodology that we presented previously which uses Stanford NER for public policy text analysis *Fig. 4*.

Topic Modeling based Methodology alter the initial Gov-DAF knowledge base building pipeline in public policy text analysis phases, as it uses Mallet LDA for topic modeling to recognize and cluster keywords *Fig. 5*.

Gov-DAF knowledge base building pipeline has two implementations as discussed above. Following is an overall presentation of the Gov-DAF knowledge base building pipeline components:

A. Approach independent Gov-DAF knowledge base building pipeline stages

1) Input Public Policy

Gov-DAF inputs or assets are the public policies required to be analyzed along with the public policy analytical aspects i.e. public policy objectives entered by the domain experts. A policy can be either an old public policy that is under analysis or a new public policy that is under discussion to be introduced. Public policy documents can be input in many document formats e.g. pdf.

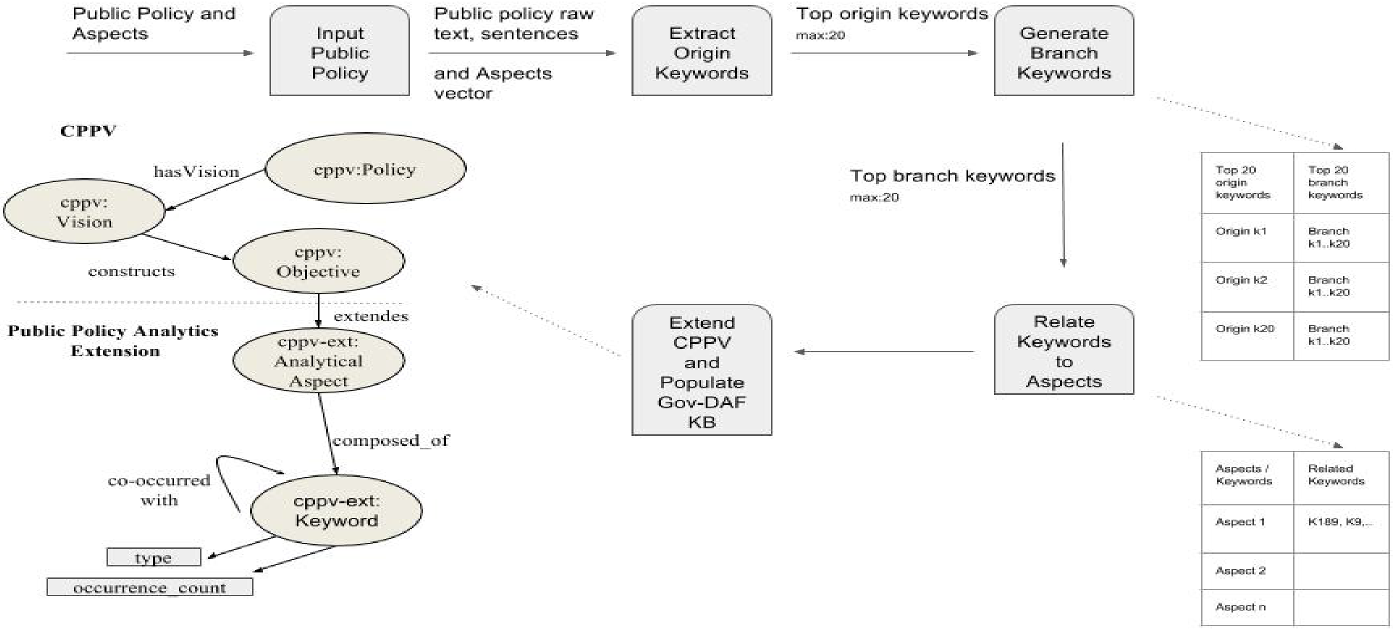
Input public policy is divided into sentences to suit Mallet LDA analysis according to the Topic Modeling Methodology. Public Policy aspects reflect the main components of a public policy. Aspects are defined and fed into the developed system by domain experts and decided upon by the user during public policy input phase. A public policy raw text, sentences, and aspects vector are the output of this phase.

2) Relate Public Policy Keywords to Aspects

Public Policy aspects are to be connected to a set of origin and branch keywords or topic clusters. This set should be strongly descriptive of the aspect. Keywords will be used for both citizen opinions collection and analysis. Automating this process is carried out using semantic similarity score approach by calculating the semantic relatedness score between public policy aspects and public policy keywords then nominating top related keywords for every aspect. This process quality is measured and reported in section 3.

3) Extend CPPV and Populate Gov-DAF Knowledge Base

The CPPV is extended with public policy analytics extension (cppv-extended) with Class (cppv-extended: Keyword). The keywords are extracted from the public policy either using NER method or Topic Modeling method. It is also extended by properties (cppv-extended:type, occurrence_count) as shown in *Fig. 3*, and *Table 2*,. Then the Gov-DAF knowledge base is populated with public policy aspects i.e. objectives defined by domain experts and Gov-



Top 20 origin keywords	Top 20 branch keywords
Origin k1	Branch k1..k20
Origin k2	Branch k1..k20
Origin k20	Branch k1..k20

Aspects / Keywords	Related Keywords
Aspect 1	K189, K9,..
Aspect 2	
Aspect n	

Fig. 4. Gov-DAF Knowledge Base Building Pipeline using Named Entity Recognition Based Methodology

DAF users.

TABLE II. “SIMPLIFIED” REPRESENTATION OF GOV-DAF KNOWLEDGE BASE IN TURTLE

cppv and cppv-ext	**Sample KB entries**
cppv:Policy	ex:BioStrategy a cppv:Policy ;

cppv:Vision	ex:vision1 a cppv:Vision;
	ex:policy1 cppv:hasVision ex:vison1;

cppv:Objective	ex:O1 a cppv:Objective;
	ex:vision1 cppv:constructs ex:o1;

cppv-ext:AnalyticalAspect	ex:AA1 a cppv-ext:AnalyticalApect;
	ex:O1 cppv-ext:extends ex:AA1;

cppv-ext:Keyword	ex:Education a cppv-ext:Keyword;
	ex:AA1 cppv-ext:composed_of ex:Education;

cppv-ext:type	ex:Education cppv-ext:type “Person”;
cppv-ext:occorance_count	ex:Education cppv-ext:occorance_count “100”;

B. Named Entity Recognition Methodology dependent pipeline stages Fig. 5

1) Extract Origin Keywords

Public policy text contains places, persons and organizations within its sentences. All these elements or entities are possible candidates for being the main public policy actors. Using Stanford NER, entities are recognized and tagged with its type for possible multidimensional correlation analysis. Filtering redundant occurrences of those keywords hold their occurrence score to use it as a significance weight of the keyword, sort keywords based on the significance weight and finally use top 20 keywords as the possible main policy keywords candidates.

2) Generate Branch Keywords

After origin keywords recognition and filtration process, a keywords network exploration process is started using DISCO library to extract distributional related words using co-occurrences. DISCO is founded over the similar words clustering algorithm introduced in [21]. The keywords network is then relaxed and expanded by nominating the top 20 related branch keywords to every origin keyword according to their semantic relatedness score calculated by DISCO library.

C. Topic Modeling Methodology dependent pipeline stages Fig. 6

1) Detect Topics Clusters

Using Mallet LDA topic modeling methodology [17] over public policy sentences extracted in phase one, will allow for an enhanced approach for Public policy text analysis and keywords extractions and clustering.

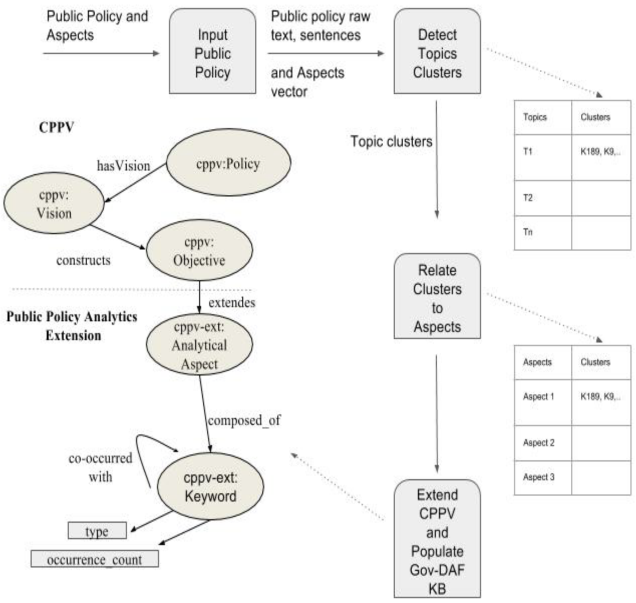


Fig. 5. Gov-DAF Knowledge Base Building Pipeline using Topic Modeling Based Methodology

III. TESTING AND RESULTS

To test Gov-DAF knowledge base building pipeline accuracy, the aforementioned methodologies were applied over two different public policy sample documents (UK Bioenergy Strategy and Irish National Plan for Equity of Access to Higher Education [22], [23]) for keywords extraction and keywords clustering around manually extracted public policy aspects i.e. objectives. The extracted keywords using both methods were related back to objectives using human evaluation to test both entity recognition accuracy of NER and Mallet LDA, and the classification accuracy of DISCO. The following performance measures used in machine learning domain [24] were calculated to assess the results *Fig 6*.

As shown in *Table 3*, and *Fig. 7*, Topic Modeling Methodology appears to prove better performance compared to Named Entity Recognition Methodology as it has higher

$$Classification\ \ Accuracy\ \ = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$Error\ Rate = \frac{(FP+FN)}{(TP+TN+FP+FN)}$$

$$Precision = \frac{TP}{(TP+FP)}$$

$$Recall = \frac{TP}{(TP+FN)} \qquad f\ measure = \frac{2\,P\,R}{(P+R)}$$

Fig. 6. Accuracy Measures Equations Adopted from [24]

Accuracy, Precision, Recall, F measure and lower Error rate. Depending on those results Gov-DAF knowledge base pipeline will adopt and enhance Topic Modeling Methodology in the coming phases of this research.

TABLE III. ACCURACY MEASURES

	NER		Topic Modeling	
	PolicyA	*PolicyB*	*PolicyA*	*PolicyB*
Accuracy	0.3	0.51	0.82	0.74
Error Rate	0.7	0.49	0.18	0.26
F measure	0	0	0.75	0.62
Precision	0	0	1	1
Recall	0	0	0.6	0.45

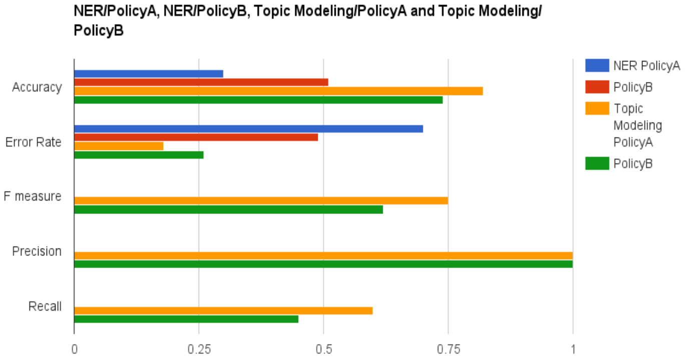


Fig. 7. Accuracy Measures Compared

IV. CONCLUSION AND FUTURE WORK

Gov-DAF was proposed in [13], [12], Gov-DAF is addressing the problem of lack of tools to support critical government decision making in which knowledge of citizen opinions expressed on social media constitute a critical input. In this paper, Gov-DAF knowledge base component was presented, named “Gov-DAF knowledge base building pipeline” and implemented. Gov-DAF Knowledge base building pipeline was implemented using two methods; Named Entity Recognition Based Methodology, and Topic Modeling Based Methodology. Both Methods implementation details are reported and accuracy measures are presented. Reckoning to the accuracy indices results, Gov-DAF will adopt and enhance the Topic Modeling Based Methodology for the next phases of Gov-DAF implementation; namely (Opinion Harvesting, Opinion Mining and Satisfaction Estimation).

REFERENCES

[1] J. M. Kelly and D. Swindell, “A multiple--indicator approach to municipal service evaluation: correlating performance measurement and citizen satisfaction across jurisdictions,” *Public Adm. Rev.*, vol. 62, no. 5, pp. 610–621, 2002.

[2] S. L. Percy, “Response time and citizen evaluation of police,” *J. Police Sci. Adm.*, vol. 8, no. 1, pp. 75–86, 1980.

[3] G. G. Van Ryzin, “Expectations, performance, and citizen satisfaction with urban services,” *J. Policy Anal. Manag.*, vol. 23, no. 3, pp. 433–448, 2004.

[4] W. G. Skogan, “Citizen satisfaction with police encounters,” *Police Q.*, vol. 8, no. 3, pp. 298–321, 2005.

[5] B. Stipak, “Citizen satisfaction with urban services: Potential misuse as a performance indicator,” *Public Adm. Rev.*, pp. 46–52, 1979.

[6] D. Swindell and J. M. Kelly, “Linking citizen satisfaction data to performance measures: A preliminary evaluation,” *Public Perform. Manag. Rev.*, pp. 30–52, 2000.

[7] E. W. Welch, C. C. Hinnant, and M. J. Moon, “Linking citizen satisfaction with e-government and trust in government,” *J. public Adm. Res. theory*, vol. 15, no. 3, pp. 371–391, 2005.

[8] J. M. Kelly, “Citizen Satisfaction and Administrative Performance Measures Is there Really a Link?,” *Urban Aff. Rev.*, vol. 38, no. 6, pp. 855–866, 2003.

[9] G. G. Ryzin, D. Muzzio, S. Immerwahr, L. Gulick, and E. Martinez, “Drivers and consequences of citizen satisfaction: An application of the American customer satisfaction index model to New York City,” *Public Adm. Rev.*, vol. 64, no. 3, pp. 331–341, 2004.

[10] S. de Walle and G. G. Van Ryzin, “The Order Of Questions In A Survey On Citizen Satisfaction With Public Services: Lessons From A Split-ballot Experiment,” *Public Adm.*, vol. 89, no. 4, pp. 1436–1450, 2011.

[11] R. Bandari, S. Asur, and B. A. Huberman, “The Pulse of News in Social Media: Forecasting Popularity,” *arXiv Prepr. arXiv1202.0332*, 2012.

[12] M. Adel Rezk, A. Ojo, G. A. El Khayat, and S. Hussein, “A Government Decision Analytics Framework Based on Citizen Opinion,” in *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*, 2016.

[13] M. Adel Rezk, A. Ojo, G. A. El Khayat, and S. Hussein, “A Proposed Government Decision Support System Based on Citizens Interactions over Social Networks,” in *Proceedings of the FIFTH International Conference on Information and Communication Technology in Our Lives 2015*, 2015.

[14] M. Adel Rezk, M. H. Aliyu, H. Bensta, and A. Ojo, “Core Public Policy Vocabulary.”

[15] The Open Knowledge Foundation, “Ckan - The open source data portal software.” [Online]. Available: http://ckan.org/. [Accessed: 09-Nov-2016].

[16] J. R. Finkel, T. Grenager, and C. Manning, “Incorporating non-local information into information extraction systems by gibbs sampling,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 363–370.

[17] A. K. McCallum, “MALLET: The Machine Learning for Language Toolkit.” [Online]. Available: http://mallet.cs.umass.edu/. [Accessed: 09-Nov-2016].

[18] S. E. Whang, D. Marmaros, and H. Garcia-Molina, “Pay-as-you-go entity resolution,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 5, pp. 1111–1124, 2013.

[19] P. Kolb, “Disco: A multilingual database of distributionally similar words,” *Proc. KONVENS-2008, Berlin*, 2008.

[20] P. Kolb, “Experiments on the difference between semantic similarity and relatedness,” in *Proceedings of the 17th Nordic Conference on Computational Linguistics-NODALIDA’09*, 2009.

[21] D. Lin, “Automatic retrieval and clustering of similar words,” in *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, 1998, pp. 768–774.

[22] U. Government, “UK Bioeneregy Strategy,” 2012.

[23] H. E. Authority, “National Plan for Equity of Access to Higher Education 2015-2019,” 2015.

[24] F. Guillet and H. J. Hamilton, *Quality measures in data mining*, vol. 43. Springer, 2007.