SOTICS 2019

The Ninth International Conference on Social Media Technologies, Communication, and Informatics

ISBN: 978-1-61208-757-3

November 24 - 28, 2019

Valencia, Spain

SOTICS 2019 Editors

Pascal Lorenz, University of Haute-Alsace, France

Nitin Agarwal, University of Arkansas at Little Rock, USA

SOTICS 2019

Forward

The Ninth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS 2019), held on November 24 - 28, 2019- Valencia, Spain,, was an event on social eco-informatics, bridging different social and informatics concepts by considering digital domains, social metrics, social applications, services, and challenges.

The systems comprising human and information features form a complex mix of social sciences and informatics concepts embraced by the so-called social eco-systems. These are interdisciplinary approaches on social phenomena supported by advanced informatics solutions. It is quit intriguing that the impact on society is little studied despite a few experiments. Recently, also Google was labeled as a company that does not contribute to brain development by instantly showing the response for a query. This is in contrast to the fact that it has been proven that not showing the definitive answer directly facilitates a learning process better. Also, studies show that e-book reading takes more times than reading a printed one. Digital libraries and deep web offer a vast spectrum of information. Large scale digital library and access-free digital libraries, as well as social networks and tools constitute challenges in terms of accessibility, trust, privacy, and user satisfaction. The current questions concern the trade-off, where our actions must focus, and how to increase the accessibility to eSocial resources.

We take here the opportunity to warmly thank all the members of the SOTICS 2019 technical program committee, as well as all of the reviewers. We also kindly thank all the authors who dedicated much of their time and effort to contribute to SOTICS 2019. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also gratefully thank the members of the SOTICS 2019 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that SOTICS 2019 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of social eco-informatics. We also hope Valencia provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

SOTICS 2019

Committee

SOTICS Steering Committee
Lasse Berntzen, University of South-Eastern Norway, Norway
Nitin Agarwal, University of Arkansas at Little Rock, USA
Andrea Nanetti, School of Art, Design, and Media | Nanyang Technological University, Singapore

SOTICS Industry/Research Advisory Committee
Roman Shtykh, Yahoo Japan Corporation, Japan
Xin Shuai, Thomson Reuters, USA
Andrea Cimino, Institute for Computational Linguistics (ILC-CNR), Pisa, Italy

SOTICS Publicity Chair
Jose Miguel Jimenez, Universitat Politecnica de Valencia, Spain // University of Haute Alsace, France

SOTICS 2019 Technical Program Committee

Najeeb G. Abdulhamid, Brunel University London, UK
Bilal Abu Salih, Curtin University, Australia
Safa'a AbuJarour, University of Potsdam, Germany
Witold Abramowicz, Poznan University of Economics and Business, Poland
Millicent Akotam Agangiba, The University of Mines and Technology, Ghana
Nitin Agarwal, University of Arkansas at Little Rock, USA
Swati Agarwal, BITS Pilani, Goa Campus, India
Ahmet Aker, University of Duisburg-Essen, Germany / University of Sheffield, UK
Esther Andrés Pérez, ISDEFE / Technical University of Madrid, Spain
Mehdi Asgarkhani, Ara Institute of Canterbury, New Zealand
Liz Bacon, University of Greenwich, UK
Grigorios N. Beligiannis, University of Patras, Greece
Gerardo Berbeglia, Melbourne Business School, Australia
Valentina Beretta, IRD Montpellier, France
Lasse Berntzen, University of South-Eastern Norway, Norway
Brian Blake, University of Arkansas at Little Rock / Harding University / Acxiom Corporation, USA
Dominik Bork, University of Vienna, Austria
Christos Bouras, University of Patras | Computer Technology Institute & Press «Diophantus», Greece
Piotr Bródka, Wrocław University of Science and Technology, Poland
María Luisa Carrió Pastor, Universitat Politècnica de València, Spain
Ilknur Celik, Cyprus International University, Northern Cyprus
Subrata Chakraborty, University of Southern Queensland, Australia
Manoj K. Chinnakotla, Artificial Intelligence and Research (AI & R), Microsoft, India
Christina Christodoulakis, University of Toronto, Canada
Andrea Cimino, Institute for Computational Linguistics (ILC-CNR), Pisa, Italy
Taiane Ritta Coelho, Federal University of Parana, Brazil

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Restrictions Towards the Adoption of Social Media Platforms by Civil Servants

Joshua Ebere Chukwuere
Department of information systems
North-West University
Mahikeng, South Africa
joshchukwuere@gmail.com

Sikedi Ramawela
Department of Information Systems
North-West University
Mahikeng, South Africa
ramawelasikedi@gmail.com

Abstract—Technology is changing continually and the introduction of Web 2.0 has brought about innovations such as social media (SM) and other inventions. Social Media brings about continual communication and engagement for the public, including civil servants. It allows civil servants (users) to create profiles, connect to existing profiles and communicate with others. It also allows civil servants (users) to view, comment, like, navigate, as well as share views and ideas with others. However, there are many challenges that hinder civil servants from adopting social media platforms (SMPs) for work and non-work related purposes, especially in developing countries. Academic scholars are widely exploring various challenges confronting individuals and organisations from adopting social media platforms, but little is known of the facts that hinder civil servants from adopting social media platforms. Therefore, this study seeks to identify possible challenges that hinder civil servants from adopting social media platforms, both for work-related and personal activities. The study deployed close-ended questionnaires involving 252 civil servants within selected cities in the North West Province, South Africa. The study found that personal customs and traditions, organizational policy, high cost of data and others cause hindrances for civil servants to adopt social media platforms in their daily lives.

Keywords-civil servant; social media (SM); social media platforms (SMPs); social media network (SMNs); web 2.0; organizations; developing countries.

I. INTRODUCTION

The use of technology and its ever-evolving innovation has affected people's daily lives through access to the Internet. Social media is seen as the 21st century innovation with a great deal of impact in public services. The use of the Internet, therefore, gave rise to social media platforms, such as Facebook, Instagram, Skype, Twitter, WhatsApp and many others. Social media platforms create seamless mediums for cost-effective and efficient communication of important information locally and globally [1]. Social media platforms are used in government and non-governmental organizations to improve communication, productivity and provide services [2]. According to Ferreira and Du Plessis [3], social media platforms improve effective communication among individuals with common goals and interests. This increase creates connectivity between organisations and employees, organisations and customers, and among employees. However, organisations perceive negative effects of social media platforms on employees' productivity and commitment to work [3]. This means that organisations see no benefit in the adoption and usage of social media platforms at work.

Notwithstanding the benefits attached to social media platforms, some organisations and civil servants in developing countries are yet to adopt it due to some challenges. However, there are few or no empirical studies in developing countries dealing with challenges, barriers and restrictions towards employees or civil servants adopting social media for either work or non-work related purposes. Therefore, this study is aimed at investigating some of the restrictions that hinder civil servants in developing countries from adopting social media platforms and many more. Ferreira and Du Plessis [3] suggest that civil servants are challenged by Internet usage and technological infrastructures in their effort to use social media platforms. Furthermore, many civil servants are digital immigrants (born or brought up before technology), which is a tall order for them to adopt 21st century technology trends and innovations, while others are digital natives (born or brought up during technology age (digital age)). The study is structured into problem statement and research objectives, literature review, research methodology, data analysis and discussions, findings and recommendations, conclusion and future research and references.

II. PROBLEM STATEMENT AND RESEARCH OBJECTIVES

The advent of social media platforms (SMPs) has defined and is still redefining historical and cultural patterns of human communication and interaction across different spheres. Pamphlets, newspapers, magazines, leaflets and brochures were the most-used communication channels within organisations before the arrival of social media [4]. Furthermore, for civil servants to exchange information and contents within and outside the organization, emails and telephones could be used. Social media platforms created an innovative 21st century method of communication and establishing connections and friendships for individuals and organisations [5]. Social media platforms allow for the construction of virtual profiles and engaging with other users. The connected individuals have differences such as social status (class), culture, ethics, values, beliefs and many others [6].

The platforms allow users to share commonalities across developed and developing countries on various issues.

 ISBN: 978-1-61208-757-3

Previously, social media was used by private sectors [7], but has become a norm for the public sectors. The platform is used in social, health and educational environments and even by government and non-government officials [2]. Ferreira and Du Plessis [3] state that Social media platforms assist humans to achieve continuous connections with known and unknown friends as well as new ones with the intention of sharing common interests and information, also acquiring new skills and knowledge. To some people, including civil servants, the continuous connectivity is among the main reasons why they use social media. Furthermore, some issues on social media platforms can trigger 'social activity', which causes low productivity, privacy and security risks, phishing, identity theft, social engineering, spam, malware and many more [3].

Social media platforms present organisations with more opportunities to engage with their customers timely, effectively and directly in many ways and get them involved in branding and feedback [8]. The platform offers low-cost engagement with its employees and other stakeholders such as customers. Civil servants use social media in many ways [9]. The researchers further suggest that social media users should be presented with the opportunity to communicate, exchange ideas and participate in social activities, engagements and discussions. It enables users to begin a new relationship, maintain existing ones and engage in business transactions. Scholars have proven great opportunities that lie behind the effective usage of Social media platforms. However, current studies have failed to capture the challenges or hindrances that stop civil servants from adopting social media platforms in their daily lives. According to Woods [7], there is a limited study on the usage of social media in government. Then, the primary objective of this study seeks to investigate some common restricting factors that limit civil servants in developing countries from adopting social media platforms.

Based on the research gaps identified, the research objective further seeks to determine whether civil servants perceive that social media platforms were designed for specific groups; is the platform time consuming, destroying cultural values, and increase security challenges. Furthermore, this study aimed to determine the challenges confronting civil servants in adopting Social media platforms.

III. LITERATURE REVIEW

This section provides an overview on existing studies covering different kinds of social media platforms, positive aspect of civil servants using social media platforms in the work setting, and the negative aspect of social media platform on civil servants. The section guides readers to understand the underpinning literature in regard to this study.

A. *Kinds of Social media platforms*

Web 2.0 powers the present social media platforms. Web 2.0 is a second generation of the World Wide Web (www), which provides individuals the ability to generate contents and share them online using the social media platforms. Web 2.0 is also called social web, participatory or participative web. The Web 2.0 platforms assist the web to be engaging. According to [2]-[10], Web 2.0 services are wikis, blogs, RSS, podcasting, social booking, microblog and many others. Each of these facilitates the functionality of social media platforms. Al-Badi [2] and Kietzmann et al. [30] believe that Social media platforms enable 'sharing', 'relationships', 'conversation', 'presence', 'reputation', 'identity' and 'group'. The categorization of Social Media is a topic of discussion by different scholars [2][10]-[15]. Overall, on estimate, there are more than 120 different types of social networking sites or applications categorized into 16 classifications, as shown in Figure 1. Social Media sites, applications or platforms are interchangeable in this study.

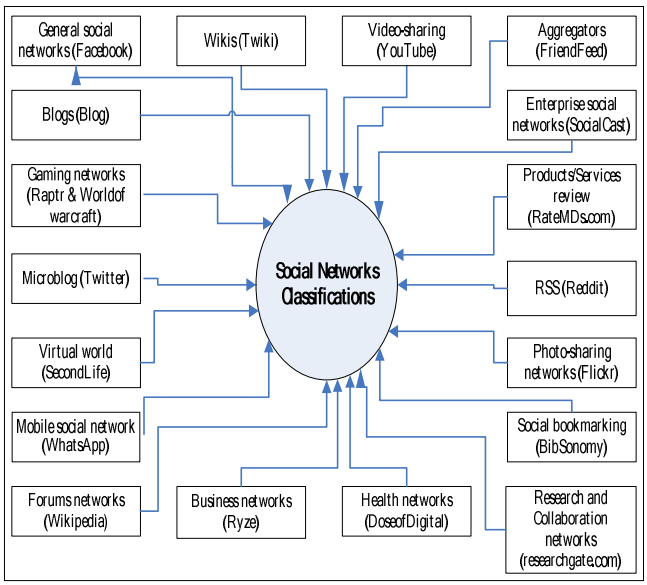

Figure 1. Social media classifications, adapted from Al-Badi [2]

Different professionals including government and non-government workers are using social networking sites or applications (Figure 1) in constructing ideas and knowledge and interacting with others on work-related and personal activities. According to Richthammer et al. [16], social media have become part of humans in exchanging information and communication with one another. The platforms have revolutionised the ways human interactions are taking place. The penetration of social Media in the public and private sectors provides positive and negative impacts to the users.

B. *Positive aspect of civil servants using social media platforms in the work setting*

Scholars are aware of the benefits of social media networks (SMNs) or social media platforms on civil servants' daily interactions [17]. Social media networks are an interlinked platform that is conceptualised to shape social groups with different forms of interactions and activities. It transfers information between enrolled participants [18].

 ISBN: 978-1-61208-757-3

Work-related benefits of social media networks allow users to connect with the external experts, making and reinforcing ties with associates, gathering quality data, and advancing information sharing with colleagues and executives [19]. A study by Zhang et al. [28] determines how social media affects employees in China in their jobs. The findings show that social media promotes positive commitment and engagement in an organisation; it increases job satisfaction and decreases turnover intention of employees [28]. In addition, Song et al. [29] carried a qualitative study proves that work-related and non-work-related social media promotes team engagement and improve performance.

Empirically, social media provides positive impacts for organisational employees whether for work-related or non-work-related purposes [28][29][30]. Social media enables individuals to air their views and make contacts online. It is not the same as the old standard media, because it presents robust and flexible connective platforms for all. According to Hysa et al. [17], Social media platforms lie in the power of Web 2.0 by allowing individuals and organisations to build contents and exchange them. It contains content that might be in the form of a video, audios, images and texts with the ability to unite societies, and help individuals build connections [20]. As a Web 2.0 platform, Social Media allows individuals to build personal or organisational virtual profiles and connect with others [2]. Cilliers et al. [21] add that it grants workers (civil servants) a direct opportunity to reach out to other colleagues and customers. It further enables civic servants to stream meetings online and discussions without leaving their workplace. The positive benefits associated to Social media platforms are huge for organisations and civil servants (employees).

Organisations currently use the influence of Social media platforms to identify new business opportunities, building new interest groups and associations, new industries, distinct knowledge, skills and proficiency. On the other hand, the platform enables organisations to store and transfer information of various marketing needs and techniques. In addition, social media helps organisations to connect with potential workers [22] and share valuable contents and information with employees. Few organisations value the usage of internal social media networks to determine and map their workers' intelligence and improve cooperate communication to empower fast access. These internal social networks will have tremendous impacts on organisational turnover [18], because employees will be able to connect with each other on work-related issues. The discussed literature proves that social media platforms enrich civil servants, corporate and personal connections.

C. *Negative aspect of social media platforms on civil servants*

Regardless of the advantages and commitment of Social media platforms on organisations and civil servants as specified above in the work environment, it also presents numerous threats, as it transforms the way individuals connect in their organisations [22]. Social media's effortless connection has the potential of creating problems for the operational, tactical and strategic employees. Hysa et al. [17] believe that social media networks can distract and lead to slow civil servants' work productivities because of large amounts of time spent on the platform. The authors also suggest that the misuse of the platform by an employee can create negative perceptions about a company and its brand. To another author, social media networks can be used obsessively in the work environment, which may result in preventing employees from working to their full capacity and potential [21].

Moreover, civil servants may download online content that consumes large amounts of bandwidth, which has proven to be costly for the user, especially in developing countries. There are also additional dangers that civil servants may encounter when using social media. These include viruses that could damage the systems in the workplace, employees accessing pornographic sites, as well as posting defamatory comments or classified data on social media sites [21]. To Leftheriotis and Giannakos [19], social media sites are regarded as a time-waster and a security ambush for civil servants. However, according to Akram and Kumar [23], negatively, the use of social media can be time consuming; mistakes can go viral and many more. According to this study, social media platforms can expose civil servants to cyberbullying, exposing organisational confidential information to the public, increasing distractions, reducing employee-to-employee contact time with each other, increasing fraud, and destroying reputations.

IV. RESEARCH METHODOLOGY, DATA ANALYSIS AND DISCUSSION

This study deployed quantitative research methodology, which collected numerical data [24]. According to Walliman [25], quantitative research analyses data using statistical and mathematical processes to reach results. As a quantitative research, the researchers used questionnaires in data gathering involving private and public organizations in Mmabatho and Mahikeng. Questionnaires was circulated to the participants on hardcopies and 252 copies were collected through random sampling techniques among civil servants, with the aim to determine basic factors that restrict them in the adoption of social media networks for both personal and work-related activities and achieving other objectives. Analysis was done on the collected data with the application of Microsoft Excel spreadsheet and Statistical Package for the Social Science (SPSS).

This section is structured into demographic data (questions) and the research objectives which are: the factors that restrict civil servants from adopting social media platforms (cultural factors hindering the adoption of Social media platforms and civil servant challenges in adopting Social media platforms), civil servant perception on the social media platforms (design; time consuming, destroys cultural values and security challenges), and Pearson's correlation. Through this structure, the collected data are analyzed to establish the data meanings and interpretation.

Inversely, the followings are the research questions: What are the factors that restrict civil servants from adopting social media platforms? What are civil servants' perceptions on the

 ISBN: 978-1-61208-757-3

social media platforms? However, the subheading below was drawn from the research objectives.

A. *Demographic data*

The primary purpose of the demographic questions was to determine the respondents' home language and educational qualifications. Both demographic questions are important to study, example, language is a good form of culture while educational qualifications were used on the Pearson's' correlation. However, majority of the participants were female civil servants.

TABLE I. HOME LANGUAGE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Afrikaans	13	5.2	5.2	5.2
	English	13	5.2	5.2	10.3
	IsiNdebele	9	3.6	3.6	13.9
	IsiXhosa	22	8.7	8.7	22.6
	IsiZulu	19	7.5	7.5	30.2
	SiSwati	10	4.0	4.0	34.1
	Southern Sotho	16	6.3	6.3	40.5
	Setswana	115	45.6	45.6	86.1
	Northern Sotho	19	7.5	7.5	93.7
	Tshivenda	8	3.2	3.2	96.8
	Xitsonga	8	3.2	3.2	100.0
	Total	252	100.0	100.0	

Furthermore, the researchers asked a question to determine the participants' language. There are eleven (11) official languages in South Africa. According to Table 1 above, 45.6% (115) speak Setswana and other languages follow. The study shows that majority of the participants are Setswana speaking people.

This study also seeks to understand the educational qualification of the respondents. It was recorded that majority of them hold higher and university educational qualifications 73.8% (186). Secondary certificate holders were 23.14% (60) while 2.4% (6) are primary school certificate holders.

B. *The factors that restrict civil servants from adopting social media platforms*

Number of factors affect or hinder civil servants from adopting social media platforms. According to Ferreira and Du Plessis [3], the followings threats can cause users' withdrawal from Social media platforms: low productivity, privacy and security risks, distraction, phishing, identity theft, social engineering, spam, malware, lack of trust and the list continues. This objective section aimed at discovering factors that may hinder the civil servants from adopting Social media platforms. The objective involves three questions to determine these factors.

Cultural factors hindering the adoption of social media platforms - This question was intended to understand the different cultural factors that might hinder civil servants from adopting social media platforms. According to Cilliers et al. [6], online users have personal and social differences such as social status (class), culture, ethics, values, beliefs and many others, which impact their decision to adopt Social media platforms.

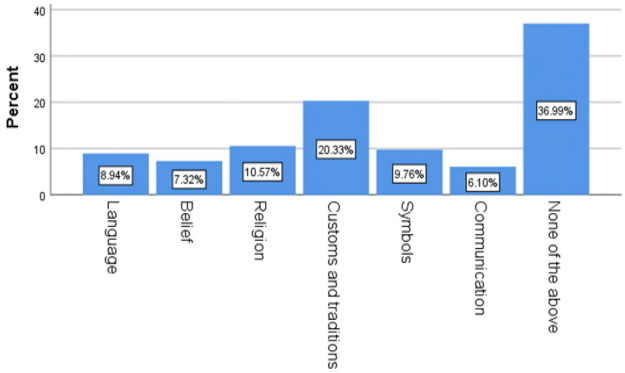

Figure 2. Cultural factors hindering the adoption of social media

Figure 2 above indicates that 36.99% (93) of the respondents believed that none of the cultural factors hinders their adoption of social media, while 20.33% (51) attribute customs and traditions as hindrance factors. The overall findings illustrate that most civil servants are not hindered by their cultural factors from adopting social media platforms; however, customs and traditions have some level of hindrance to the users.

Workplace policy influences civil servants' usage of social media platforms in working environments - This question seeks to find out whether workplace policy prevents civil servants from using Social media platforms at work. The findings revealed that 37% (94) state affirmatively that their workplace policy hinders workers from adopting or using social media at work. Importantly, this response contradicts the findings made by [22], which state that organizations use social media to engage with workers and join cooperative communication. Cilliers et al. [21] further added that Social Media enables employees to reach out to other colleagues and potential customers. Moreover, 36.9% (93) of the respondents indicated that workplace policy does not deny them the use of social media, while 14.2% (36) were uncertain, and 11% (28) responded that none of the options was applicable to them. The findings proved that workplace policy restricts civil servants from adopting social media platforms in their work environment.

Civil servant challenges in adopting social media platforms - There are a number of challenges that users face when adopting social media platforms. Table 3 seeks to determine those challenges.

 ISBN: 978-1-61208-757-3

TABLE II. CHALLENGES IN ADOPTING SOCIAL MEDIA PLATFORMS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	High cost of data	103	40.9	41.9	41.9
	Lack of internet connection	19	7.5	7.7	49.6
	Lack of internet enabled smartphone, laptop, desktop, tablet	11	4.4	4.5	54.1
	Lack of knowledge on the benefits of social media	25	9.9	10.2	64.2
	Personal, religious and cultural beliefs	40	15.9	16.3	80.5
	None of the above	48	19.0	19.5	100.0
	Total	246	97.6	100.0	
Missing	System	6	2.4		
Total		252	100.0		

Table 2 above demonstrates that 40.9% (103) are affected by the high cost of data and other challenges. The results clearly show that the high cost of data is the most common hindrance confronting most participants.

C. Civil servants perception on the social media platforms design; time consuming, destroys cultural values and security challenges

Civil servants' perception on the design and the time spent on social media platforms determines whether it will be used on a work-related basis or for personal activities. The impact of social media platforms on civil servants' culture will certainly determine whether they will accept the usage. Furthermore, security situations will have an impact. The questions below are aimed at determining all these from the civil servants.

TABLE III. SOCIAL MEDIA PLATFORMS DESIGNED FOR SPECIFIC GROUPS; TIME CONSUMING, DESTROYS CULTURAL VALUES AND SECURITY CHALLENGES

	Questions	Options	Frequency	Percentage
1	Social media platforms designed for specific group	Strongly agree	55	21.82
		Neutral	38	15.1
		Strongly disagree	154	61.1
2	Social media is time consuming	Strongly agree	162	64.28
		Neutral	57	22.6
		Strongly disagree	29	11.50
3	Social media destroys cultural values	Strongly agree	91	36.1
		Neutral	84	33.3
		Strongly disagree	73	28.96
4	Social media usage exposes one to online security challenges	Strongly agree	143	56.7
		Neutral	72	28.6
		Strongly disagree	37	14.68

Table 3 comprises of four questions with the aim to understand whether civil servants perceived that Social media platforms was designed for specific group of individuals, and time consuming, destroying cultural values and exposing users to security threats and challenges. Question 1 proves that social media platforms are not designed for specific group of individuals, with 61.1% (154) in support. The findings clearly show that civil servants believe that social media is not designed for ordinary people and not for everyone to use.

The second question shows that 64.28% (162) strongly suggest that social media usage is time consuming. In support of the findings, Leftheriotis and Giannakos [19] noted that employees view Social Media as a time-waster. In addition, Cilliers et al. [21] also indicated that employees could become overly obsessed with the use of social media, which may prevent them from performing their professional duties effectively. According to the third question, 36.1% (91) of the respondents stated that the use of social media destroys their cultural values. The study proves that civil servants' usage of social media platforms gradually destroys their cultural values and norms, while 56.7% (143) of the respondents suggested that the use of social media exposes users to different forms of online security challenges.

Leftheriotis and Giannakos [19] indicated that others view Social Media as a security threat, and therefore support the findings of this study. Moreover, Cilliers et al. [21] explained the dangers that employees may come across from the usage of social media, such as downloading online viruses that could harm systems or devices, as well as posting confidential data and information on social media.

D. Pearson's correlation

Table 4 above presents the questions involved in the correlation and the assigned abbreviations to each of them.

TABLE IV. RESEARCH QUESTIONS AND ABBREVIATIONS

	Research question	Abbreviation
1	Do you think cultural factors are considered while designing social media platform?	B1
2	Do you think culture destroys social media usage?	B2
3	I think that the usage of social media will expose me to online security challenges	B3
4	I think that the usage of social media will destroy my cultural values	B4
5	I think that the usage of social media is time consuming	B5
6	I think that social media is not designed for ordinary people to use	B6
7	What challenges are you facing in adopting social media?	B7
8	Does the workplace policy deny the use of social media in the working environment?	B8
9	Do these cultural factors hinder you from adopting social media?	B9

The correlations are aimed at determining whether there is a relationship between the variables [26][27]. It was used in this study to determine the strength between two variables that lie between +1 and -1. The p-value is the center point for this study, which means that when the p-value is less than or equal (⩽) to 0.5, then the relationship is significant to the study to make predictions [26]. This interpretation means that any variables above p-value (0.5) will be rejected and excluded from the study. Pearson's coefficient scale as applied in the study: Weak uphill or downhill ± 0.30, Moderate uphill or downhill ± 0.50, Strong uphill or downhill ± 0.70, Strong downhill -1, Perfect = 1 and No linear relationship = 0 [26] [27]. Table 5's contents were

 ISBN: 978-1-61208-757-3

based on **. Correlation is significant at the 0.01 level (2-tailed) and *. Correlation is significant at the 0.05 level (2-tailed).

TABLE V. PEARSON'S CORRELATIONS FOR THE STUDY

Research question	Correlation range	P-value	Correlation	Level of significance
Level of education/B1	0.167**	0.008	Weak uphill	Significant
Level of education/B4	0.127*	0.045	Weak uphill	Significant
B1/B2	0.272**	0.000	Weak uphill	Significant
B1/B4	0.185**	0.003	Weak uphill	Significant
B1/B7	-0.147*	0.021	Weak downhill	Significant
B2/B4	0.235**	0.000	Weak uphill	Significant
B3/B9	0.142*	0.026	Weak uphill	Significant
B4/B5	0.345**	0.000	Weak uphill	Significant
B4/B6	0.314**	0.000	Weak uphill	Significant
B4/B9	0.178**	0.005	Weak uphill	Significant
B5/B6	0.167**	0.009	Weak uphill	Significant
B6/B7	0.278**	0.000	Weak uphill	Significant
B6/B9	0.226**	0.000	Weak uphill	Significant
B7/B8	-0.162	0.011	Weak downhill	Significant
B7/B9	0.171	0.007	Weak uphill	Significant

According to Table 5, the findings show that there is a weak uphill (positive) relationship between participants' level of education and their perception of whether cultural factors were considered while designing social media platforms (B1). It means that the more educated participants are, the more they think cultural factors are considered, and vice versa. The study also presents a weak uphill (positive) relationship between the level of education of civil servants and B4. It indicates that the more participants' education levels increase, the more they believed that the use of Social Media would destroy their cultural values and beliefs, and vice versa.

There is a weak uphill (positive) relationship between B2 and B1; this finding indicates that the more participants believe that social media destroys their culture; the more they believe that their cultural factors are not considered while designing social media platforms. Furthermore, participants believe that an increase in the perception of B1 means an increase in participants' perception of B4 and this is a weak uphill (positive) relationship. In addition, there is a weak downhill (negative) relationship between B1 and B7. The findings show that a decrease in participants' perception of B1 can lead to a decrease in B7, and vice versa.

There is a weak uphill (positive) relationship between B2 and B4. The findings indicate that an increase in B2 will automatically lead to an increase in B4, and vice versa. B2 and B6 are positively (weak uphill) related to each other, which implies that upward movement on any of the variables will draw the other on the same movement. This means that the participants' perception of B2 will positively influence B6. Furthermore, the study found that B3 has a weak uphill (positive) association with B4; the positive association suggests that an increase in any of the variables will move another to increase. The study found that participants believed that the usage of social media would expose them to online security challenges, which will positively impact on how the usage of social media destroys their cultural values.

Questions B3 and B5 are moderately uphill (positively) linked; this finding proves that B3 will influence B5 to increase, and vice versa. This means that an increase in the perception of civil servants of B3 will also increase their perception of B5. B3 and B9 have a weak uphill (positive) relationship with each other, meaning that an increase in one will also cause the other variable to move in the same direction. According to the relationship between B4 and B5, there is a weak uphill (positive) association between both variables. This kind of relationship shows that both variables will always move in an upright position. The study also proves that B4 and B6 are in a weak uphill direction, which means that both variables increase together and influence each other.

The statistical association between question B4 and B9 indicates a weak uphill relationship, which indicates that when participants' perceptions of B4 increases, then B9 will increase too. The finding proves that the way participants think of social media to destroy their cultural values, and increase how they feel about different cultural factors that hinder them from adopting social media. While B5 and B6 are positively related, having weak uphill relationships, it simply means that both variables are moving in an upward direction at the same time. As they believe that social media is time consuming, it increases, so their perception is that social media is not designed for ordinary people to use. According to the study, B6 and B7 are weak uphill related, which shows that both variables have a good and positive impact on each other. Both variables move in an upward direction and influence each other.

The study found that B6 and B9 have weak uphill impacts on each other, which indicates that the increasing views of participants on B6 will lead to their perception of B9 to also increase. This shows that participants think that social media was not designed for ordinary people to use, which increases as their views of B9 increase. Again, the association between B7 and B8 is viewed as weak downhill (negative); the nature of the correlation implies that when the participants' view of B7 decreases, the outcome will automatically affect B8 to move in the same direction. In addition, B7 and B9 have a weak uphill impact on each other. The outcome shows that a positive perception of participants on B7 will also affect B9 positively (weak uphill movement).

V. FINDINGS AND RECOMMENDATIONS

To this study, the adoption of social media platforms presents opportunities as well as threats to user's civil servants. Little is known in South Africa on the opportunities and threats of social media platforms on civil servants both on personal and work related purposes. This study has closed the existing research gap with a new insight and discovering of knowledge and ideas. Hysa et al. [17] suggest that scholars are aware of the benefits of social media platform on civil servants' daily interactions with colleagues as well as families and friends. Here are some of the key findings from the study, which will help employees, and academicians in developing countries to understand what restricts employees adopting social media:

 ISBN: 978-1-61208-757-3

- The overall findings illustrate that most civil servants are not hindered by their cultural factors from adopting social media platforms; however, customs and traditions have some level of hindrance for the users.
- The findings prove that workplace policy restricts civil servants from adopting social media platforms in their work environments.
- The results clearly show that the high cost of data is the most common hindrance confronting most participants. However, there are other challenges, as shown in Table 3.
- The findings clearly show that civil servants believe that social media is designed for ordinary people and for everyone to use.
- The study further proves that civil servants' usage of social media platforms gradually destroys their cultural values and norms.
- The respondents suggest that the use of Social media platforms exposes users to different forms of online security challenges and threats, which act as a hindrance in their adoption.
- In the end, it was found that the educational level of civil servants would determine whether they would believe that their cultural factors were considered in the design of social media.

End-users expectations on social media platforms keep changing especially in the developing countries where cultural, social, economic and political (CSEP) challenges abound. The study recommends that for these restrictions to be conquered, social media platform designs should be based on civil servants' (users') cultural attributes. Furthermore, social media platforms should give users the ability to customize the platform to suit their culture. There should be enough training for the civil servants to understand the effective usage of social media platforms. Organizations should implement social media policies to promote effective and productive usage. Civil servants should undergo training on such policies. The cost of data for internet access should be reduced significantly to promote productive usage among different ages, populations and classes. Furthermore, civil servants should avoid borrowing social trends on social media that disvalue their own cultures and values. Managing different forms of bridges and exposing of personal information means adequate security, such as anti-virus, firewall, encryptions and many more across all social media platforms.

VI. CONCLUSION AND FUTURE RESEARCH

There are different forms of restrictions that challenge civil servants in adopting social media platforms in developing countries. This study was able to highlight those challenges and restrictions confronting the civil servants' intentions with the ability to inform organisations and designers alike. The future for social media platforms is bright in developing countries; researchers should carry out research to determine what civil servants are really doing on social media when the opportunity arises. Research should be conducted to understand the factors that push civil servants to adopt social media. This kind of study has to be conducted on the influence social media on different population groups.

REFERENCES

[1] M. K. Abubakar, M. N. Patricia, O. O. Samuel, O. O and A. Totolo, "Factors affecting adoption of Social Media by women's non-governmental organisations (WNGOs)", International Journal of Library and Information Science, vol. 9, no. 9, pp. 96-106, 2017

[2] A. H. Al-Badi, "The adoption of Social Media in government agencies: Gulf Cooperation Council case study", Journal of Technology Research, vol. 5, no. 1, 2014.

[3] A. Ferreira and T. Du Plessis. "Effect of online social networking on employee productivity", South African Journal of Information Management, vol. 11, no. 1, pp, 1-11, 2009.

[4] K. Pillay and M. S. Maharaj, "Social Media and mobile communications adoption patterns of South African civil society organisations", South African Journal of Information Management, vol. 16, no. 1, pp, 1-8, 2014.

[5] C. Kaitlin, "Social Media Changing Social Interactions", Student journal of media literacy education, vol. 1, no. 1, 1-17, 2010.

[6] L. M. Lekhanya, "Cultural influence on the diffusion and adoption of Social Media technologies by entrepreneurs in rural South Africa", International Business & Economics Research Journal, vol. 12.12, vol. 12, 2013.

[7] W. E. Woods, "Government 2.5: The Impact of Social Media on Public Sector Accessibility", 2016.

[8] B. Sago, "Factors Influencing Social Media Adoption and Frequency of Use: An Examination of Facebook, Twitter, Pinterest and Google+", vol. 3, no. 1, pp. 1-14, 2013.

[9] M. M. Zwiers, "Exploration and explanation of the multiple ways of how employees make use of Social Media in their daily practices" (Master's thesis, University of Twente), 2014.

[10] M. Shrivastava, T. Paperwala and K. Dave, "Trends in web technologies: Web 1.0 to Web 3.0 & beyond", In The International Information Systems Conference (iiSC) 2011 Sultan Qaboos University, Muscat, Sultanate of Oman, pp, 73, 2011.

[11] D. Nicholas and I, Rowlands, "Social Media use in the research workflow", Information Services & Use, vol. 31, no. 1-2, pp, 61-83, 2011.

[12] P. Andersen, "What is Web 2.0?: ideas, technologies and implications for education", Bristol: JISC, vol. 1, no. 1, pp. 1-64, 2007.

[13] M. J. Culnan, P. J. McHugh and J. I. Zubillaga, J. I, "How large US companies can use Twitter and other Social Media to gain business value", MIS Quarterly Executive, vol. 9, no. 4, 2010.

[14] A. Kassel, "Social networking: A research tool", FUMSI article, 2018.

[15] L, Safko, "The Social Media bible: tactics, tools, and strategies for business success", John Wiley & Sons, 2010.

[16] C. Richthammer, M. Netter, M. Riesner, J. Sänger and G. Pernul, "Taxonomy of social network data types", EURASIP Journal on Information Security, vol. 1, no. 11, 2014.

[17] B. Hysa, A. Mularczyk and I. Zdonek, "Social media–the challenges and the future direction of the recruitment process in HRM area", Studia Ekonomiczne, vol. 234, pp. 54-67, 2015.

 ISBN: 978-1-61208-757-3

[18] E. Tokarcíková, ”Influence of social networking for enterprise´ s activities”, Periodica polytechnica social and Management Sciences, vol. 19, no. 1, pp. 37-41, 2011.

[19] I. Leftheriotis and M. N. Giannakos, “Using Social Media for work: Losing your time or improving your work?”, Computers in Human Behavior, vol. 31, pp. 134-142, 2014.

[20] M. Drahošová and P. Balco, “The analysis of advantages and disadvantages of use of Social Media in European Union. Procedia Computer Science, vol. 109, pp. 1005-1009, 2017.

[21] L. Cilliers, W. T. Chinyamurindi and K. Viljoen, “Factors influencing the intention to use Social Media for work-related purposes at a South African higher education institution”, SA Journal of Human Resource Management, vol. 15, no. 1, pp. 1-8, 2017.

[22] B. Aguenza, A. H. Al-Kassem and A. M. Som, “Social Media and productivity in the workplace: Challenges and constraints”, Interdisciplinary Journal of Research in Business, vol. 2, no. 2, pp. 22-26, 2012.

[23] W. Akram, and R. Kumar, “A Study on Positive and Negative Effects of Social Media on Society”, International Journal of Computer Sciences and Engineering, 5, 347-354, 2017.

[24] D. Muijs, “Doing quantitative research in education with SPSS”, Sage, 2010.

[25] N. Walliman, “Research methods: The basics”, Routledge, 2017.

[26] J. E. Chukwuere, “Towards a culture-oriented e-learning system development framework in higher education institutions in South Africa”, (Doctoral dissertation, North West University, South Africa), 2017.

[27] D. J. Higgins, “The Radical Statistician: A Beginners Guide to Unleashing the Power of Applied Statistics in the Real World”, (5th Ed). California: Jim Higgins Publishing, 2015.

[28] X. Zhang, L. Ma, B. Xu, B and F. Xu, “How Social Media usage affects employees’ job satisfaction and turnover intention: An empirical study in China”. Information & Management, 56(6), 103136, 2019.

[29] Q. Song, Y. Wang, Y. Chen, J. Benitez and J. Hu, “Impact of the usage of Social Media in the workplace on team and employee performance”. Information & Management, 2019.

[30] J. H. Kietzmann, K. Hermkens, I. P. McCarthy and B. S. Silvestre, “Social media? Get serious! Understanding the functional building blocks of social media”. Business horizons, 54(3), 241-251, 2011.

 ISBN: 978-1-61208-757-3

Smart Contacts as APIs

Athanasios Priftis
Information Systems Department
HESSO / HEG-GE,
Geneva, Switzerland
e-mail: athanasios.priftis@hesge.ch

Joël Israel
Information Systems Department
HESSO / HEG-GE,
Geneva, Switzerland
e-mail: joel.israel@hesge.ch

Jean-Philippe Trabichet
Information Systems Department
HESSO / HEG-GE,
Geneva, Switzerland
e-mail: jean-philippe.trabichet@hesge.ch

Abstract— **Although blockchain protocols have existed for some time now, a focused analysis on smart contracts, as Application Programming Interfaces (APIs) for user driven and web based applications, is clearly missing. APIs as abstract interfaces can inspire us in designing smart contract based applications and information infrastructures. Such an approach has an impact both on the architecture and coding of applications. In this article, we will use our pilot on managing building rights within the City of Geneva to demonstrate how the architecture, design and implementation of smart contracts can be advanced. Initiating the creation of new applications and services based on the smart contracts characteristics, such as forced temporality and immutability and transparency, comes with new opportunities and challenges. Blockchain could be more than an innovative technology, a building block of new forms of social applications and infrastructures through the design of smart contracts as APIs.**

Keyword-Smart contracts; Blockchain pilot; APIs; Web APIs; information infrastrucure and application; user-generated content.

I. INTRODUCTION

The main purpose of our article is to examine a blockchain - smart contract infrastructure, inspired by APIs, in a real life pilot. This research and application effort, launched late 2018, is still in progress within the Geneva City administration. We will start the article presenting smart contracts in relation to the concepts of information infrastructures, in particular Application Programming Interfaces (API). An analysis will follow, presenting the work that is taking place, involving various actors: a research group of the University of Applied Sciences in Geneva, collaborating with several public administration departments of the Geneva State in the area of building rights management and house development, notably the Cantonal Office of Housing and Urban Planning (DALE) and selected private entities. The goal of this cross-organizational, action oriented, research effort is to co-produce a set of smart contracts, developed as APIs, facilitating the open and transparent execution of urban planning processes, while designing a multi-stakeholder governance infrastructure of smart contracts. This information infrastructure could set the basis for initiating hybrid, public – private, services in the future. Finally, we will discuss the importance of coding smart contracts as APIs. We will make appear some crucial characteristics of smart contracts as key elements both in the area of building rights management and the smart contracts' themselves.

This is how our paper is structured. In section II, we describe in more detail what a standard API is and discuss the sociotechnical aspect of information infrastructures, mainly in terms of public governance. In section III, we present how smart contracts create applications and spaces of social decision. Finally, in sections IV and V, we describe the, API driven, architecture of our information infrastructure, related to our pilot application, and discuss further work and challenges.

II. UNDERSTANDING APIS AS INFORMATION INFRASTRUCTURES

As already demonstrated in previous research efforts [1], we cannot rely on the modern disciplinary methods and frameworks of knowledge in order to think and interpret the transformative effect which new technology is having on our culture. It is precisely these methods and frameworks that modern technology requires us to rethink. Smart contracts as APIs can intersect the current state of opacity in application development and contribute to our understanding of semantic rules to user created applications.

An API can be understood as an abstract interface establishing parameters for computational exchange. These parameters can be accessed and incorporated for the creation of any number of possible interfaces. In other words, it acts as an interface, mainly by representing and defining the possible functions of the exposed information elements, in the form of tools that express, and make available, certain functions of these elements. In this way, an API creates a standardized method to facilitate forms of exchange between various information elements and computational agents to make them interoperable and independent of their respective implementations [2]. As mentioned, this is done through an API's establishing of specified procedures, typically through establishing parameters of access through the assigning of various identifiers, priorities and restrictions that can be operated upon within API-facilitated exchanges.

 ISBN: 978-1-61208-757-3

In regards to which information elements can and cannot be surfaced and shared, an API can be seen as both an entry point into the black box of a particular computational service, but also as a clearly defined possibility towards other possible exchanges with this service. At present, the term API refers specifically to the category of APIs known as Web-based APIs. A Web API encapsulates and specifies all of the valid messages that two or more computational parties can request and accept while communicating via network protocols [3].

Bucher [4] provides important understandings to the sociotechnical questions at play in API practices. They pertain to API-supported fields such as application ecosystems and social media platforms. They are better understood when combining insights from fields, such as software studies with ethnographic approaches into how developers produce and make sense of code in their work with APIs. The importance of APIs as both practical connective enablers and abstract infrastructures for networked computational practices is a key element of our analysis. By focusing on smart contracts as an API implementation in information infrastructures, we aim to give a few suggestions for how anyone working with them might think openness and terms of inclusivity set upon practices of sharing, participation and exchange.

Information infrastructures are closely linked to social innovation. They are considered as a significant part of Information and Communication Technologies (ICT) innovations, the development and study of which comprises both the technological components, as well as, the social aspects. The analysis of these information infrastructures includes technological characteristics, capabilities, interactions and negotiations between actors involved in their development. Information infrastructures have to be developed through a collaborative approach, as actors have to give up control some over their data and systems to realize mutual benefits, supported by governance mechanisms making this possible. The entire setting in which actors operate may change because of a social innovation [5]. This requires organizations to develop advanced social and collaborative capabilities, to be able to realize new modes of public governance. Social factors affect the development, adoption, change, operations, and stability of information infrastructures, as well as, the application and services linked to them [6].

In this context, learning from APIs while developing smart contracts, can be extremely useful in the following areas: a) designing and deploying the overall architecture of our application, b) understanding and explaining, the unique possibility of each smart contract as an API, serving a larger, user oriented information infrastructure and c) establishing user driven parameters negotiating the relationship between transparency, openness, business model and integration of systems. There areas are answering to questions such as what data is closed, what is open, what is made accessible, what is kept internal to a system, what is open to edition and how this possibility to edit is, actually, taking place.

Our approach provides more experience and results at this exact point. Smart contracts as APIs become an important element to give some sense on how data is being circulated, made accessible and inaccessible. Even more, they allow us to (re)think, and at times intervene, to the overall rules of governance of platforms and applications around us.

III. ON SMART CONTRACTS

Understanding smart contracts as applications and spaces of social decision making, needs a more detailed analysis. This is what this section attempts to do.

A. Smart Contracts and their design as applications

The term smart contract was introduced by Nick Szabo back in 1997 [7]. Smart contracts are self-executing computer programs that implement a set of functionalities. They are based on business rules and contractual agreements. Smart contracts, very much like APIs, can automate business logic by embedding, verifying, and enforcing the contractual clauses of an agreement without intervention from intermediaries. The main characteristics of smart contracts include machine readability and distributed code running on a blockchain platform. Smart contracts, similarly to APIs, can be part of an application program, but can also act autonomously for a predefined period distributed [8].

Blockchain technology established the ground for the implementation of smart contracts as pieces of code that consist of executable functions and state variables. Specifically the execution of a function changes the state of the variables according to related logic implementation. Nowadays, the Ethereum blockchain protocol [9], is the most widely used technological platform for the development of the smart contracts, using Solidity, an object oriented high-level language, as the implementation language.

The design of a smart contract consists of their conceptual and technical part. The latter requires the setup of the blockchain nodes, the definition of the business functions, the description of the processes between the users and the application template design for the definition of the smart contract. The conceptual design consists of the description and classification of business rules that will be extracted from information carriers (e.g., documents or code). The specification of conversion of extracted rules to smart contract functionalities using domain knowledge, formally represented as ontologies [10]. The extracted information gains semantic meaning from the exploitation and usage of standardized knowledge representation, such as ontologies and semantic rules. Adopting semantic rules incorporated into, and enforced by a smart contract, can be facilitated by using smart contract templates. The templates can serve as the skeleton for generating the final smart contract to be used in the blockchain network.

B. On Contracts, Smart Contracts and Social Decision making

As Dupont and Maurer argue, blockchain technologies differ from traditional social systems that validate, maintain and enforce contracts between people (e.g., accountancy and legal systems), because crypto-contracts tend to build social and functional properties within the system [11]. In other

 ISBN: 978-1-61208-757-3

words, where lawyers and judges are needed to enforce legal regulations and notaries are needed to validate certain legally binding contracts, the blockchain allows for the validation of smart contracts and their enforcement in its own right without the necessity for arbitrating third parties. This implies that in contrast with conventional contract laws, which are necessarily coupled with their human validators and enforcers, blockchain technologies are capable of establishing and maintaining forms of political organization that are (at least in the virtual realm) self-sustaining [12].

The decentralized enforcement of smart contracts "dematerializes" or rather depersonalizes the auditing authority: it eradicates the need for human arbitrators such as notaries or accountants. While traditional contracts can be described as textually expressed voluntary agreements between two or more contracting parties that require human arbitration to be validated, audited and enforced, a smart contract appears as a mechanism that can be made binding by means of computational scrutiny, without human interference. However, the work of contracting remains embedded in social interactions, namely the act of consenting to a specific contractual reality. The aspects that are delegated to the technology are the validation, storing and enforcement of the contractual clauses.

As an initial governance method for our case study, linked to smart contract and information infrastructure challenges, we envisage a consensus-driven approach. As Klievink and Janssen note, the consensus process is well suited for a society where technological and economic progress is within reach [13]. This approach has two clear advantages over existing alternatives. First, the initiation of an ongoing discussion with interested parties, as to achieve an early alignment on governance rules, mainly through a proposal and voting possibilities. Second, pilot participants gradually develop strong incentives to resolve conflicts early in the process to secure the viability of the application.

IV. PILOT: OPEN REGISTERS FOR BUILDING RIGHTS

In this section, we will examine the context of our pilot and describe the, API driven, architecture of our information infrastructure.

A. Context

The DALE (Office cantonal du Logement et de la Planification Foncière, State of Geneva), in collaboration with the University of Applied Sciences in Geneva (HEG-GE) are co-producing a public register with set of smart contracts facilitating the open and transparent execution of urban planning processes. This existing initiative attempts to: a) test and authenticate the execution of a process between several entities of the domain, through the application of smart contracts and b) make proposals around a multi-stakeholder governance of smart contracts for public services.

A recent study by Credit Suisse [14] describes the challenging situation around an "affordable" housing-to-buy in Geneva, highlighting, the urbanism consequence: transportation, pollution, environment, moving of population to other countries - areas. Thus, new solutions, services and policies around building rights and urban planning are becoming urgent. The existing informational infrastructure includes a detailed analysis of the business process around building permissions. The overall view of how the building permission is processed today, includes: a) public administration actors - departments' participation and roles, b) decision process and status of a building permit and c) rules of when and how are building rights are calculated.

There are few selected information and consultation initiatives, driven by the public administration, with professional and local populations before the opening of new building zone. This process is set in order to generate building rights in specific areas with some kind of citizen participation. Evaluating this situation, we concluded that the process is largely opaque to the outside and confined within the public administration actors. At the same time, opening a new building zone can last up to four years, depending on various circumstances. Moreover, the implementation process of the accepted projects, in a specific building area, is not available to the public. These initial elements justify the main goal of this pilot: managing the building rights process in a more open, educated and collective way.

B. An API driven architecture

The architecture of our application tries to serve the need for more openness, transparency and collective management by the following core socio-technical elements: a) decentralized, permissive and editable storage of all data collected within our application, b) easy and transparent smart contract deployment and scrutiny for the related administrative processes and c) possibility for users to create proposals and vote for the proposals of others.

We are using the InterPlanetary File System's (IPFS) API to interact with it as our storage system. IPFS is a protocol and network designed to create a content-addressable, peer-to-peer method of storing and sharing hypermedia in a distributed file system [15]. Similar to a torrent, IPFS allows users to not only receive but host and edit content.

As opposed to a centrally located server IPFS is built around a decentralized system of user-operators who hold a portion of the overall data, creating a resilient system of file storage and sharing. The main method is the following: a hash is obtained based on the image file's binary codes. The file is retrieved by searching for it with its hash. It is not possible to replace an image, with another one, because the file is changing when its hash changes. The hash code is immutable on the Ethereum Blockchain and the file is immutable on IPFS.

For our application to interact with IPFS, we are using the method, presented in Figure 1. An IPFS node dials to other application instances using WebRTC:

 ISBN: 978-1-61208-757-3

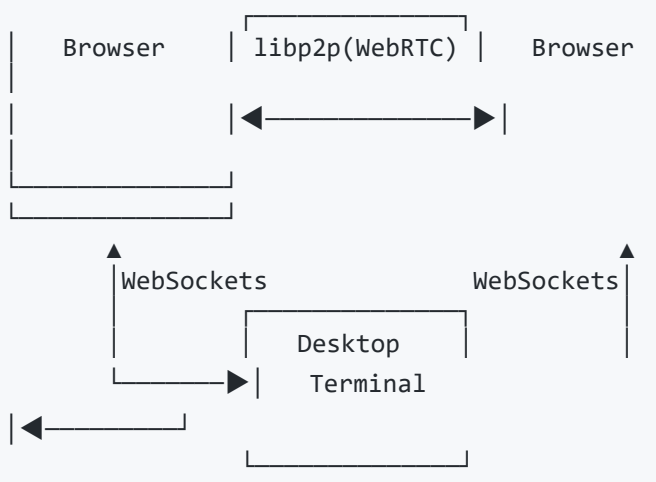

Figure 1. Application diagram with an IPFS node that dials to other instances using WebRTC

We are using the Nuxt.js framework, ideal for creating Vue.js applications and abstracting away the client/server distribution. The two important functions deployed are the ADD function (which takes a message ready to be posted and returns the information of the created hash) and the READ function read (which takes a hash and returns the message created).

The next step is to add the discussion with Ethereum, here the logic of the mechanism. We first check if we have local data to be taken into consideration through the user's browser: if no, we collect the data by reading the smart contract, pointing us to the file that needs to be recovered. If yes, we check with the smart contract that the referenced data are still valid, we update the smart contract and notify all the users that this change took place, followed by the online documentation. This process is facilitated by an appropriate middleware known as MetaMask. MetaMask is an application acting as a bridge that allows to run Ethereum apps from the user's browser without running a full Ethereum node.

The result of these architectural choices is an early prototype, published in June 2019, validated as a minimum viable product from the actors of the pilot. It follows the architecture described above and gives access to read and edit an initial building rights' table (DAB) describing the number of building rights allocated to a specific construction area, as well as, the exact parts of land associated to them. In Figure 2, we present a view of the main page of the DAB prototype:

Figure 2. The DAB prototype accessible at https://proto3.ynternet.org/

This initial table describing building rights in a specific construction area is at the core of the case study. It is the main register allowing for building applicants (site developers, architects, private entities) to interact with the public administration and claim their rights to execute a building project in a selected area. Below follows a more detailed view of this, now, decentralised and blockchain validated table. Figure 3 demonstrates how the editing of the initial building rights table in the DAB is presented:

Tableau initial des DABs partagé

Tableau validé

Tableau de répartition des droits à bâtir			Localisation des droits à bâtir		
Parcelle N°	Surface Parcelle en ZD3	DAB	BAT a1, a2	BAT b1, b2, b3	BAT c1
1128	3125	150	0	0	0
2458	3560	3300	0	0	1800
2566	3420	3040	450	2590	0
3456	2050	2328	0	0	0
3879	2122	1980	540	0	640
4321	2345	2000	0	2000	0
TOTAL	16622	15315	990	4590	3000

Figure 3. Editing the initial building rights table in the DAB prototype accessible at https://proto3.ynternet.org/

Regarding proposal creation of the application, our architecture implements the following general idea: when a user wants to send a new data through a smart contract, or even create a new one, we add his/her proposal on a stack of proposals with a dedicated function (addProposal). When we accept a proposal, we empty the stack and replace the data with the one in the proposal (acceptProposal that calls updateHash). This point brings us to the semantic data structure of the application. As already mentioned, IPFS allow us to post whatever type of data we wish, raw data or encrypted data. The application is now saving data in the JavaScript Object Notation (JSON), with a history of all edits made upon this data. The final version of the application will generate several JSON data models based on different processes, for example the table of distribution of user's building- rights in the application: one can create a specific field data, mapped as an object of which each key is another object, as demonstrated in Figure 4:

```
{
 "data": {
  "parcelle_id_1": {
   "bat_1": "dab_1",
   "bat_2": "dab_2",
  }
  "parcelle_id_2": {
   "bat_2": "dab_4",
  }
 }
 "created_at": "xxxx",
 "created_by": "xxxx",
 "previousVersion": "xxxx"
}
```

Figure 4. Data structure for the distribution of user's building rights

 ISBN: 978-1-61208-757-3

This modular and decentralized architecture, described in summary above, is thought out itself as an API. It acts as an interface for semantic data to be stored and then executed, as part of, or new smart contracts. These data are organised and can be accessed through a central smart contract, an oracle, which will serve as a registry table for other smart contracts. The oracle keeps a key - value of an identifier to the address of a smart contract tracing back to all data related to it. This process is called oraclize and provides a way to get outside data from any API onto the blockchain. This point allows us to proceed to the next area and examine the actual coding structure of smart contracts as APIs.

C. Coding Smart Contracts as APIs

Coding Smart Contracts as APIs allows us to design and deploy them, with the four characteristics described in the following paragraphs. The first one consists of making public a register with its data and add a read and write function available to its users. Figure 5 details how this code operates.

```
// This is the object structure representing a record
    struct record {
      address created_by;
      address updated_by;
      address smartContractAddress;
      bool exists;
    }
    // the mapping representing the register
    mapping(string => record) internal register;
    // Public function that write a record into the register
    function write(string memory _identifier, address
_smartContractAddress) public returns (bool) {
        address creator = msg.sender;
        if (register[_identifier].exists) {
            creator = register[_identifier].created_by;
        }
        register[_identifier] = record({smartContractAddress:
_smartContractAddress, exists: true, updated_by: msg.sender, created_by:
creator});
        return true;
    }
    // Public function that read a record value from the register
    function read(string memory _identifier) public view returns
(address) {
        require(register[_identifier].exists, "This record is empty");

        return register[_identifier].smartContractAddress;
    }
```

Figure 5. Creation of a blockchain register with a read and write functions

The second characteristic is linked to the modularity smart contracts as APIs, particularly for assigning of various identifiers, priorities and rules. This includes adding the canUpdateExistingRecord parameter, set with the smart contract deployment. This parameter is stating if an existing record can be updated. In Figure 6, we present the initiation a smart contract as an API.

```
// Can you update an existing record ?
    bool internal canUpdateExistingRecord;
  // At Smart Contract deployment you must say if an existing record can
be updated
    constructor (bool _canUpdateExistingRecord) public {
      canUpdateExistingRecord = _canUpdateExistingRecord;
    }
    ...
    function write(string memory _identifier, address
_smartContractAddress) public returns (bool) {
      // Before writing into the register, check whether you are about to
update an existing record and if you have the right to do so => otherwise we
send an exception stating "Existing records can't be updated"
      require(!register[_identifier].exists || canUpdateExistingRecord,
"Existing records can't be updated");
      ...
    }
```

Figure 6. Initiating a smart contract as an API

The third characteristic is about initiating a continuous interoperability between, user-driven, applications. This is initiated by adding an event, emitted every time someone writes on the register. The data of this event are describing its full internal process and are open and reusable to all blockchain users. In Figure 7, we demonstrate the code creating events allowing for more interoperability.

```
// event that can be transmitted and followed on the blockchain
    event writeRegister(
      string _identifier,
      address _smartContractAddress
    );
    ...
    function write(string memory _identifier, address
_smartContractAddress) public returns (bool) {
      ...
      // when we write into the register, we emit the event
      emit writeRegister(_identifier, _smartContractAddress);
      ...
    }
```

Figure 7. Creating events allowing for more interoperability

The final characteristic regards the possibility for a collaborative edition of the smart contracts based on transparent rules and constant user driven improvement. This take place by adding a function that changes the value of the canUpdateExistingRecord parameter. The code presented in Figure 8 introduces a propositions' mechanism for all users of the application and will be, at a later stage of this pilot, associated to a voting function.

```
  // an event to be sent when someone change the
canUpdateExistingRecord value
    event updateCanUpdateExistingRecord(
      bool _oldValue,
      bool _newValue,
    );
    // You can modify the canUpdateExistingRecord value
    function setCanUpdateExistingRecord(bool
_canUpdateExistingRecord) public {
      emit updateCanUpdateExistingRecord(canUpdateExistingRecord,
_canUpdateExistingRecord);
      canUpdateExistingRecord = _canUpdateExistingRecord;
    }
```

Figure 8. Towards user driven collaboration and improvement

As with all public smart contracts, the code presented above can be traced online with all of its actual transactions [16]. The demonstrated characteristics showed how a smart contract designed as an API, can be implemented in a modular and open way.

 ISBN: 978-1-61208-757-3

V. CONCLUSIONS AND FUTURE WORK

Application development, in a blockchain and smart contracts context, seems like a novel opportunity to create platforms with less opacity and great collaboration for their participants. However, blockchain(s) remain protocol(s) and as we already learned in the last twenty, or so, years, decentralized protocols are neither participative, nor more democratic by default: control and centralization can take place in various levels and spheres [17]. Through this article, we tried to use APIs as an overall concept both of the architecture and the smart contracts of our pilot. Particularly, the function the oracle is cutting through existing APIs and smart contracts as applications.

Developing an API culture is a prerequisite to use any information infrastructure, particularly the ones allowing for the deployment of smart contract enabled applications. Smart contracts' immutability and forced temporality are crucial. At the time of execution of an application, there are elements that demand explicit attention and negotiation with involved stakeholders. Moreover, they need to be designed and thought our as APIs, exposing for the very beginning their objectives, rules of operation and governance.

The early experience from our building rights' management pilot is teaching us that, smart contracts adoption and administration, need to be tightly linked to specific skills within the responsible organizations. The main driver for public sector blockchain pilot initiatives, like our Geneva based pilot described in this article, is mostly based in the principles of transparency and efficiency, particularly in business process and data monitoring. Blockchain and smart contracts are unique elements for information infrastructures serving such principles.

ACKNOWLEDGMENT

This paper is possible thanks to the SCODES research project: an applied research project on blockchain protocols and smart contracts, involving five Hautes Ecoles from French-speaking Switzerland (University of Applied Sciences HES-SO). Its goal is to develop knowledge within this particular field studying and developing practical cases of use and transferring the acquired knowledge to the regional economic actors.

REFERENCES

[1] A. Priftis and J.-P. Trabichet, "The CoWaBoo protocol and applications: towards the learnable social semantic web" International Journal on Advances in Software, vol. 11, pp. 6-17, 2018.

[2] T. Espinha, A. Zaidman and H.-G. Gross, "Web API growing pains: Loosely coupled yet strongly tied," Journal of Systems and Software, vol. 100, pp. 27–43, 2015.

[3] Wikipedia. *Application programming interface*, [Online]. Available from: https://en.wikipedia.org/wiki/Application_programming_interface, [retrieved: September 2019].

[4] T. Bucher, "Objects of intense feeling: The case of the Twitter API," Computational Culture, number 3, 2013. Available from: http://computationalculture.net/objects-of-intense-feeling-the-case-of-the-twitter-api/ [retrieved: September 2019].

[5] D. Rüede and K. Lurtz, K., "Mapping the various meanings of social innovation: Towards a differentiated understanding of an emerging concept" EBS Business School Research Paper Series 12–03, Oestrich-Winkel, 2012.

[6] P. Constantinides, "The development and consequences of new information infrastructures: the case of mashup platforms", Media, Culture & Society, vol. 34(5), pp. 606-622, 2012.

[7] N. Szabo, "Formalizing and Securing Relationships on Public Networks". First Monday, vol. 2, no. 9, 1997, doi: http://dx.doi.org/10.5210/fm.v2i9.548 [retrieved: September 2019].

[8] R. O'Shields, "Smart Contracts: Legal Agreements For The Blockchain," 21 N.C. Banking Inst. pp. 180-181, 2017.

[9] V. Buterin, V. "A next-generation smart contract and decentralized application platform." Ethereum 1–36 (2014) [Online]. Available from: http://buyxpr.com/build/pdfs/EthereumWhitePaper.pdf [retrieved: September 2019].

[10] M. Wöhrer and U. Zdun, "Smart contracts: Security patternsin the ethereum ecosystem and solidity," in1st InternationalWorkshop on Blockchain Oriented Software Engineering@ SANER 2018, 2018. [Online]. Available from: https://eprints.cs.univie.ac.at/5433/7/sanerws18iwbosemain-id1-p-380f58e-35576-preprint.pdf [retrieved: September 2019].

[11] Q. Dupont and B. Maurer, B "Ledgers and Law in the Blockchain." Kings Review [Online]. Available from: http://kingsreview.co.uk/magazine/blog/2015/06/23/ledgers-andlaw-in-the-blockchain/ [retrieved: September 2019].

[12] A. Wright and P. De Filippi, "Decentralized Blockchain Technology and the Rise of Lex Cryptographia." Social Science Research Network 2580664 (2015). Available from: http://papers.ssrn.com/abstract=2580664 [retrieved: September 2019].

[13] B. Klievink and M. Janssen, M., "Developing multi-layer information infrastructures: Advancing social innovation through public–private governance". Information Systems Management, 31(3), pp. 240-249, 2014.

[14] Investment Solutions & Products Swiss Economics, "Location, location, floor plan for the Swiss Real Estate Market in 2019" [Online]. Available from: https://www.credit-suisse.com/media/assets/private-banking/docs/ch/privatkunden/eigenheim-finanzieren/swiss-real-estate-market-2019.pdf [retrieved: September 2019].

[15] J. Benet, "IPFS - Content Addressed, Versioned, P2P File System (DRAFT 3)" [Online]. Available from: https://ipfs.io/ipfs/QmV9tSDx9UiPeWExXEeH6aoDvmihvx6jD5eLb4jbTaKGps [retrieved: September 2019].

[16] Smart Contracts as API: the Oracle. Full code accessible at https://ropsten.etherscan.io/address/0xf083a44e157b9ac4b7dea543bd67547ecf57d00a#code [retrieved: September 2019].

[17] A. Galloway, Protocol: How Control Exists After Decentralization, MIT Press, 2004

Text-based Causality Modeling with Emotional Information Embedded in Hierarchic Topic Structure

Takurou Ogawa
College of Sustainable System Sciences,
School of Knowledge and Information Systems
Osaka Prefecture University
Japan
e-mail: saa01052@edu.osakafu-u.ac.jp

Ryosuke Saga
Department of Sustainable System Sciences
Graduate School of Humanities and Sustainable Systems
Osaka Prefecture University
Japan
e-mail: saga@cs.osakafu-u.ac.jp

***Abstract*—Service evaluation depends on various factors, such as assurance, responsiveness, and tangibles. Given that emotional satisfaction affects service satisfaction, analyzing both the evaluation and emotions is important in improving service. Previous studies have identified the evaluation factor and determined the degree of influence on the resulting evaluation. However, there is little effective analysis that reflects the influence of such a factor on emotion. In this study, we use hierarchal Latent Dirichlet Allocation and structural equation modeling (SEM) to express the causality relationships of service evaluation visually and quantitatively. Emotion obtained quantitatively by using sentiment analysis is newly applied to SEM to obtain knowledge reflecting the influence of emotion. As a result of the experiment, we can identify the causality of service and determine the influence of the evaluation factor and emotion quantitatively.**

Keywords-sentiment analysis; service analysis; SEM; hLDA; causal analysis

I. INTRODUCTION

In recent years, the service industry has grown rapidly such that in developed countries, there are so many markets that account for 60% to 70% of a country's gross domestic product (GDP). In the United States where GDP is the highest, the service industry's GDP is $ 15.52 trillion, accounting for 80% of the total GDP [1]. In addition, with the spread of smartphones, apps for various services (e.g., Twitter, navigation), the introduction of recommended hotels, and the rise of electronic services (e.g., Internet shopping) are rapidly increasing. With this background, the importance of services has grown in recent years. Service improvement is important as services are produced and consumed at the same time compared with products that are released and finished. Thus, analyzing the evaluation of the service in order to improve such service is important.

Service evaluation depends on various factors, such as assurance, responsiveness, and tangibles. For example, SERVQUAL evaluates the quality of service [2] with five-dimensional indicators, and Airport Service Quality (ASQ) [3] defines airport evaluation factors. As there are many factors in the evaluation of services, it is necessary to find out the evaluation factors to analyze the evaluation.

Generally, analyzing services is difficult because these have special features like Intangible, Heterogeneous, Inseparable, and Perishable (IHIP). However, there are several clues to analyze the services from the data (e.g., questionnaire). Especially, user review is useful because the review describes user experience of and perceived from the services. Therefore, it is possible to analyze the quality of service and the evaluation of service. Meanwhile, emotional satisfaction is also regarded as an important and attractive factor in service satisfaction. That is, customers experience different positive and negative emotions related to service, and these emotions influence service satisfaction [4]. Of course, these factors influence service evaluation and the emotions related to the service are implied in the user review; however, there is no study to identify and analyze evaluation factors together with emotional information.

This paper describes the method by which to perform causality analysis from text data, such as user review. In order to treat causal analysis, we use the topic-based approaches by applying a topic model to the review. In addition, the emotions for evaluation factors in the text are quantitatively determined using sentiment analysis technology. By applying topic and emotion information to structural equation modeling (SEM), we analyze the influence of each factor quantitatively.

The first contribution of this paper is that it obtains the knowledge reflecting emotional information from the user review by using sentiment analysis. Second, it understands the influence on the emotion of the evaluation factor based on the idea that emotions are essential for service evaluation factor analysis. By using SEM with path diagram, we can also analyze and understand the causality relationships among topics and their emotions associated with topics that are visually and quantitatively express.

Section 2 refers to the existing related research, section 3 explains the core technology of the analysis process, and section 4 describes analysis experiments using actual data.

II. LITERATURE REVIEW

In related research, SERVQUAL [2] measures the quality of service by measuring the gap between advance expectation and subsequent experience using five indicators prepared in advance. SURVPERF [5] measures the quality of service based on the subsequent experience alone. Related researches include a study that further increased the dimension from these five dimensions [6] and another that changed the dimension to measure the quality of electronic service [7].

 ISBN: 978-1-61208-757-3

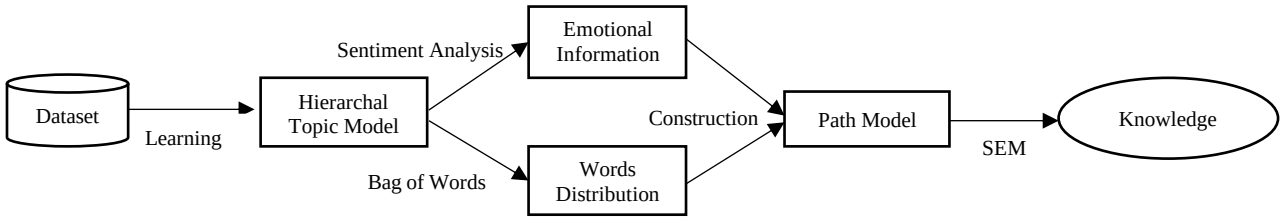

Figure 1. Analysis process

Asesthere are many evaluation indicators, it is difficult to measure all services by one standard because there are many types of services and their characteristics largely differ.

Meanwhile, related works on SEM include a study that has found relationships between customer loyalty and service quality [8] and another that has proposed a model to infer the purchase factor of the game by combining hierarchal Latent Dirichlet Allocation (hLDA) and SEM [9]. Another study made improvements to the SERVQUAL index and analyzed it with SEM [6]. These works, however, do not consider the emotion of the text.

Meanwhile, emotional satisfaction is largely believed to affect service satisfaction [4]. In relation to this, sentimental analysis is useful in comprehending and handling the emotional information. A study utilizes sentiment analysis and Latent Dirichlet Allocation (LDA) to evaluate the quality of airport services [10], while another determines the user's evaluation for each attribute by combining Airport Council International (ACI)-defined airport service quality attributes and sentiment analysis [11]. In these studies, emotion is considered one of the important factors in sales of services; thus it is essential to consider emotion. However, no study has proposed structural equation modeling that considers the emotion contained in text.

Therefore, the current paper proposes the model for SEM with emotion information. By using this model, we can acquire knowledge including emotion information visually.

III. METHODOLOGY

In this paper, the analysis is performed according to the process of Figure 1. First, topics are extracted by learning a topic model. Next, we find the emotion and topic distribution for that topic. Finally, a model is constructed based on these data and this is then analyzed by SEM so that can gain knowledge.

A. Topic Model

The topic model is a technology that tries to clarify the structure of a document group by inferring words contained in the topic based on the premise that the document group has a specific topic. In a topic model, a document is a collection of words probabilistically generated by the topic to which it belongs.

Topic models include different methods, such as latent semantic analysis (LSA) [12], LDA [13] and hLDA [14]. The LDA assumes a multi-topic model in which the document is based on mixed topics. LDA has a 1:*n* relationship between documents and topics, not 1:1 like LSA. LDA is considered to be a more natural model in documents, such as review texts that are written in one document about various aspects [13].

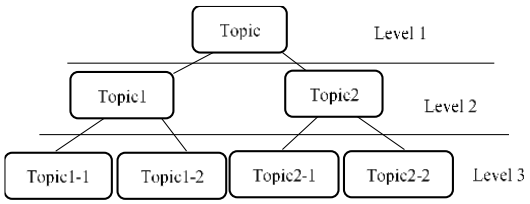

Figure 2. Hierarchal structure of topics

HLDA is an extended method of LDA and is a hierarchal model as shown in Figure 2. It has the property of automatically constructing relationships among hierarchical topics. As a learning result, a hierarchical model constructed hierarchically and a keyword group constituting each topic are generated together with their generation probabilities. The specific content of the topic can be inferred from the keyword groups of a topic. In this study, hLDA is used because it is a natural document model and the relationships between topics are defined automatically.

B. Sentiment Analysis

Sentiment analysis literally refers to the analysis of emotions. By using sentiment analysis, such as posted comments, one can determine whether consumers have negative or positive emotions and the strength of such emotions. Sentiment analysis can be performed on a per-document or per-sentence basis.

To embed emotion to SEM explained later, we have to recognize emotions on each topic for each review. In this study, we regard the average of emotion values ranging between -1(negative) and 1(positive) as document emotions by calculating Equation (1) as

$$E_{im} = \frac{1}{|T_i(S_m)|} \sum_{s \in T_i(S_m)} E(s), \qquad (1)$$

where E_{im} is the emotion about the topic T_i of the review R_m; S_m is a set of sentences in R_m and $|\ |$ is the element number of a set; $T_i(S_m)$ represents the sentence set of S_m, including the topic I; and the function E recognizes the emotion of a sentence. If there is no sentence related to a topic, (1) calculates 0 and regards this sentiment about the topic as neutral. The longer the review, the more likely it is to include other topics. Therefore, it is possible to extract emotions related to topics more accurately by focusing only on sentences containing topics in reviews.

Here, valence aware dictionary for sentiment reasoning (VADER) [15] is used as function E in the equation. This

 ISBN: 978-1-61208-757-3

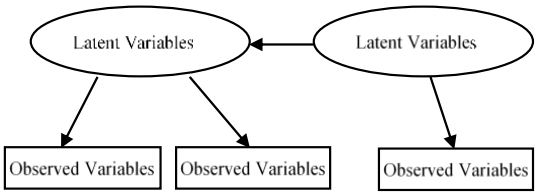

Figure 3. Path model of SEM

method is particularly accurate for sentiment analysis in social media. There are several studies that used VADER. One study analyzed the correlation of positive and negative user reviews of mobile apps before and after app update, respectively, by using VADER because VADER has the high precision in the social media field [16]. In VADER, the value of emotion is represented by -1 to 1 (the closer to -1 the more negative and the closer to 1 the more positive the emotion). Therefore, the E_{im} outputs the value between -1 and 1.

C. Strucuture Equation Modeling (SEM)

SEM [17] is a technology characterized by the use of factor analysis and regression analysis. Factor analysis is the idea that observed variables are based on some hidden factor, and the influence of the factor is to be determined by "correlation" (variance / covariance). Regression analysis is a technique for finding the relationship between a variable to be predicted (target variable) and a variable (explanatory variable, independent variable) that describes the target variable. In other words, SEM can be considered as a factor regression analysis.

The SEM can express causal relationships between variables visually and quantitatively by using a path model, as shown in Figure 3. A path model consists of three elements: latent variables, observed variables, and paths. Latent variables are factors that cannot be observed in actual. Observation variables can actually be observed and are essential for estimating a latent variable. In the path model. Latent variables are represented by ellipses and observation variables are represented by rectangles. The causal relationship between such items is represented by the path of the arrow, and the degree of influence is represented by the path coefficient.

D. Construct Path Model and Find Knowledge

Topics that cannot be observed directly are considered as latent variables serving as correspondence between SEM and topic model. The keywords that make up the topic, the emotion for the topics, and the rating values of each review are the observation variables. From the idea of the topic model that words are generated by topics, each topic is regarded as a factor and the paths from the topics are drawn to the keywords to which the topics are related. Moreover, the paths between topics are drawn from the upper topics to the lower ones based on the idea of the hierarchical structure of the hLDA topics.

Next, we explain the process of incorporating emotional information into the path model. Emotional information influences the intention of a model. Thus, we have to carefully determine how to incorporate emotional information. Generally, emotions for service are generated as perceived experience (after the service) or the expectation (before using the service). Therefore, the model is expressed by drawing a path to emotional information from each topic. When we draw a path from the topic to emotional information, the causal relationship between the emotion and the topic becomes clear. Moreover, rating evaluation is considered to be generated from the top-level topic that includes all elements. Therefore, by drawing the path from the top-level topic to the rating evaluation, the model can represent the causal relationship with the rating.

Furthermore, by comparing the values of path coefficients from the higher topics to the lower topics, it is possible to find an important factor for the rating. By paying attention to the path coefficient from the lowest topic to the keyword, we can find the degree of influence of more detailed factors. The path coefficients from each topic to emotion are large and the causal relationship with emotion could be expressed. By comparing the path coefficient from each topic to emotion, topics with a larger causal relationship with emotion can be found.

However, the path model of SEM is usually prone to model identification failure, especially if there are too many latent variables. Conversely, if the number of latent variables is less, the amount of information in the model may be too small for interpretation. As the topic is a latent variable in the path model, the number of topics must also be adjusted. We also need to remove unreliable paths and observation variables with relatively small influence.

IV. EXPERIMENTS

The purpose of this experiment is to confirm the feasibility of proposed approaches described in Section III.

TABLE I. DATA AND RESULT

Dataset Name	# of Reviews	GFI	AGFI	RMSEA	BIC
Hotel	8104	0.9025	0.8881	0.05525	9188
Airport	13444	0.9152	0.9005	0.05266	12950
App	5442	0.8979	0.8835	0.05960	6848
e-Commerce	19354	0.9213	0.9060	0.05446	19272

 ISBN: 978-1-61208-757-3

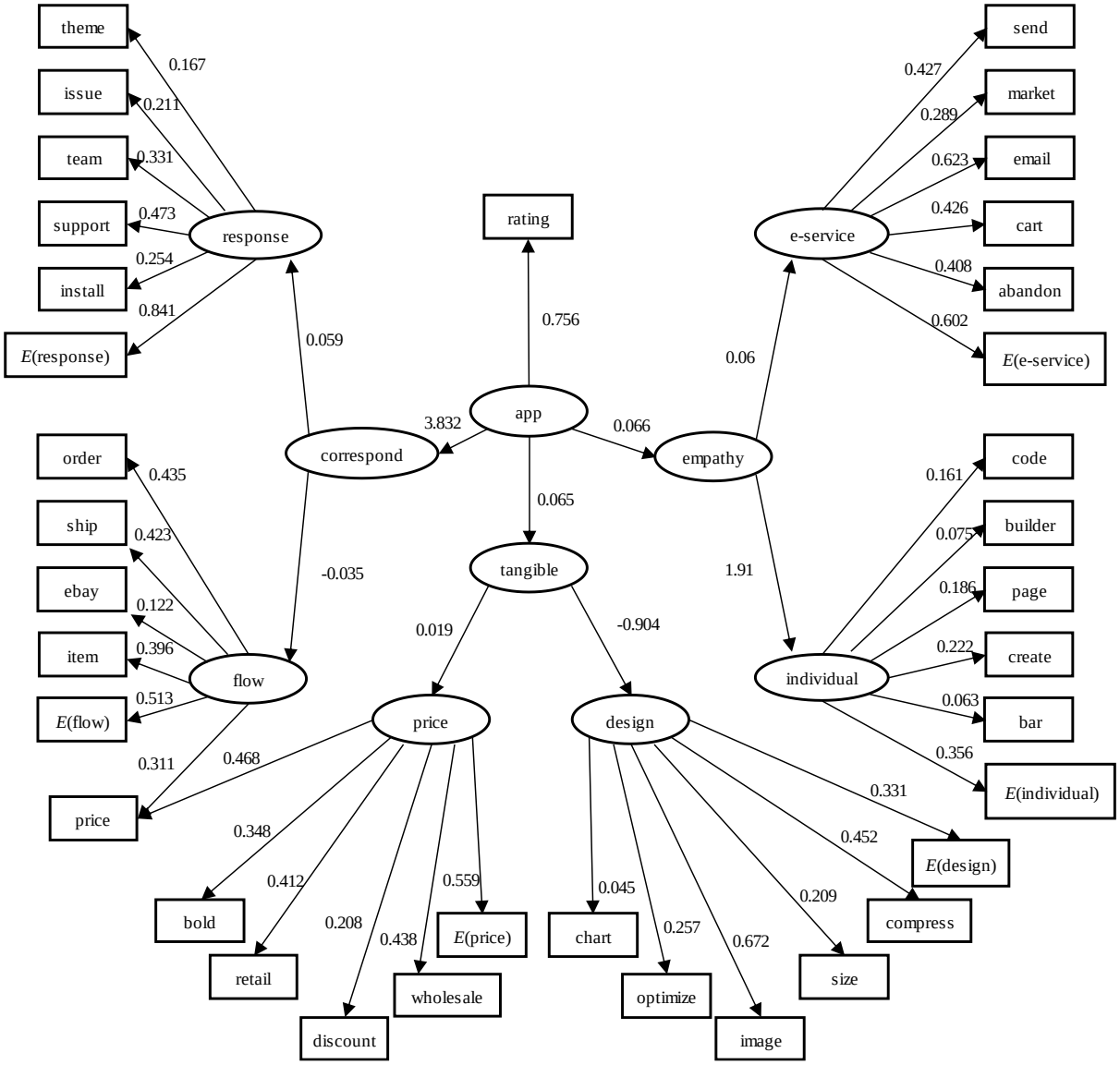

Figure 4. Analysis result of the app dataset

Furthermore, we consider the experimental results.

A. Dataset, Parameters, and Processing

In this analysis, the data must have text data and numerical evaluation data, and it is ideal to have as many review data as possible in order to apply the topic model. In addition, in order to characterize statistical data based on the concept of Bag of Words, the text of one review data must include many words. In this experiment, we employ user-reviews of the datasets published online by Kaggle and Github: the airport, hotel, app for shops and electronic services for purchasing clothes. Airport, app and electronic services reviews are collected by web scraping. Hotel reviews are provided by Datafiniti's Business Database. Each review has review text with a rating between 1 and 5 or 1 and 10. We also regard a review text as a document. In order to ensure that the topics and the appearance frequency of the feature words described are included in each document, only documents stated with more than 30 words are used. The app analyzes information from randomly extracted data. The number of reviews after these pre-processing is shown in Table 1. In this experiment, emotions on topics in the lowest level are determined for the construction of a path model. Moreover, T_i in (1) indicates a topic of the lowest level (i.e., topic in third level). Whether a sentence includes or does not include a topic is determined based on whether or not a keyword constituting the topic is included.

As criteria to evaluate the result, we use goodness of fit index (GFI), adjusted GFI (AGFI), root means square error of approximation (RMSEA), and Bayes information criterion (BIC) were used. As equations for GFI, AGFI, RMSEA, BIC,

$$GFI = \frac{tr((\sum(\hat{\theta})^{-1}(S - \sum(\hat{\theta})))^2)}{tr((\sum(\hat{\theta})^{-1}S)^2)}, \qquad (2)$$

where $\Sigma(\hat{\theta})$ is the estimated value of covariance matrix and

S is value of the actual sample covariance matrix. $tr((A)^2)$ expresses $tr(AA')$,

$$AGFI = 1 - \frac{n(n+1)}{2df}(1 - GFI), \quad (3)$$

where n is the number of observed variables and df is degrees of freedom,

$$RMSEA = \sqrt{\frac{\max\left[\frac{\chi^2 - df}{N-1}, 0\right]}{df}}, \quad (4)$$

where N is the number of samples,

$$BIC = \chi^2 - df \log(N). \quad (5)$$

And as an equation to calculate degrees of freedom,

$$df = \frac{1}{2}n(n+1) - p, \quad (6)$$

where p is the number of variables in equation. Equation (2) expresses how well the total variance in the saturation model can be explained by the estimation model. A value between 0 and 1 is taken and the closer a value is to 1, the better the model becomes. A value of 0.9 or higher is desirable. GFI is unconditionally improved in fitness as model degrees of freedom decreases. Equation (3) corrects the shortcomings of GFI and penalizes models with many parameters and high complexity. The same value as GFI is taken, and the closer it is to 1 the better the resulting model. If the model is not complex, GFI and AGFI will be close values. Equation (4) is an index that expresses the difference between the model distribution and the true distribution. The fit is good with a value of 0.05 or less, and the fit is bad with 0.1 or more. Equation (5) estimates the posterior probability based on chi-square value when the model is selected. This is used to evaluate the balance between model suitability and the amount of information and is used in carrying out relative evaluation. It is better for the value to be smaller.

In this experiment, we used several packages and libraries: Mallet package [18] for hLDA, Python's nltk package with VADER technology [19] for sentiment analysis, and SEM package of R [20] for SEM analysis.

B. Result

Table 1 shows the calculation results of the evaluation indexes for each data and analyzed models. From Table 1, we could find that hotel, airport, and e-commerce models have a GFI of over 0.9 and AGFI maintains high levels. Moreover, none of the models have an RMSEA of less than 0.05 but the values are close to 0.05. It can be said that all of models fit well to the dataset and the constructed models are reliable from the viewpoint of these indices.

As an example, let us show the result of the app dataset in Figure 4. The words at the bottom of the model are those that make up the identified topics from the text data of the review using the topic extraction with hLDA. Here, the topics (latent variables) are estimated by authors from the words that make up each topic. For example, "response" is estimated because it has a large causal relationship with "support" and is considered to be a topic related to responses to actions, such as "install," "team," and "issue." We were able to create a path model based on the hierarchical structure of a text data document group revealed by hLDA. Further, causal relationships can be analyzed by paying attention to arrow and values calculated by SEM between topics or between topics and words at the bottom of the model.

We focus on the "correspond" area with a large path coefficient from the top topic because this "correspond" can be considered as a topic having a large influence to the evaluation (rating). The "response" is also considered to be an important factor for evaluation because this path has a larger path coefficient after comparing between the two topics under "correspond." Here, the path between the latent variable "response" and the value of the emotion has a large coefficient, implying that "response" influences the emotion strongly. Given that "correspond" has a strong path, therefore, it can be considered that the emotion of response also leads to evaluation.

In the same way, when we check the other paths to emotions, we could find the relationships with and influences to evaluation. From the figure, "response," "flow," "price," and "e-service" have an effect of emotions (the paths over 0.5) and the "design" and "individual" did not. We are not certain whether the results agree or not, but this specific one indicates which topics lead to emotional satisfaction. In this way, it is possible to improve the service by quantitatively understanding the specific service factors that influence to the emotions and evaluation.

We summarize the following findings from the experiments:

- We obtained knowledge by analyzing service while considering emotions.
- We determined the impact on the rating of each topic.
- We obtained the causal relationship between each topic and emotion quantitatively and provided clues for further analyses.

V. CONCLUSION

In this paper, we analyzed the causal relationships in service by using SEM and emotional information. We constructed the path model by using hLDA and sentiment analysis between topics and emotions. The findings of the experiment using the user reviews of airports, hotels, shopping apps, and electronic services show the feasibility of our proposed model.

We also performed service analysis considering emotion and obtained knowledge reflecting emotional information from the user reviews. The consideration of emotional information is essential for service analysis, and the creation of path models with emotional information is considered effective in extracting information that helps increase service satisfaction. It is suggested that the analysis process in this paper may provide useful knowledge for service analysis and service improvement. On the one hand, this can be used by service providers in improving services and creating new services. Service providers can quantitatively find factors that have major impacts on the evaluation of services and

 ISBN: 978-1-61208-757-3

customer emotions. On the other hand, it can be used by service users to efficiently grasp the outline of services that are not formed. Although we analyzed the indefinite service in the experiments, it can be applied to other things like tangible products. The potential applicability is high because analysis is performed from the text.

As future works, we have to consider three points: emotion expression. Firstly, we extracted emotion information of topics based on (1), but this equation does not consider the length of the sentence. Nevertheless, it enables us to accurately determine the emotion on the topic by considering the weight based on the sentence length. For example, longer sentences are more likely to include other topics. Therefore, it may be possible to extract emotions related topics more accurately by reducing the impact of such sentences on emotions of specific topics. Secondly, when two or more topics are included in one sentence, even if it is used in a contrasting sentence, such as "(Text about TOPIC A) but (Text about TOPIC B)," the same emotion value is calculated for the topic. If there is a conjunction (e.g., "but"), a more accurate emotion analysis can be performed by further processing, such as dividing. Finally, in this paper, the accuracy improvement and knowledge are obtained by constructing path models under different assumptions during the construction of the path model.

REFERENCES

[1] Central Intelligence Agency: *The World Factbook: GDP - Composition, by sector of origin*. [Online]. Available from: https://www.cia.gov/library/publications/the-world-factbook/fields/214.html, [retrieved: October, 2019].

[2] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality", Journal of Retailing, vol. 64, No. 1, pp. 12-40, 1988.

[3] Airports Council International. ACI: *Airport Service Quality, ASQ. : The ASQ Barometer*. [Online]. Available from: https://aci.aero/customer-experience-asq/services/asq-barometer/, [retrieved: October, 2019].

[4] V. Liljander and T. Strandvik, "Emotions in service satisfaction", International Journal of Service Industry Management, vol. 8, Issue 2, pp. 148-169, 1997.

[5] J. J. Cronin and S. A. Taylor, "Measuring Service Quality: A Reexamination and Extension", Journal of Marketing, vol. 56, No. 3, pp. 55-68, July 1992.

[6] M. Ali and S. A. Raza, "Service quality perception and customer satisfaction in Islamic banks of Pakistan: the modified SERVQUAL model", Total Quality Management & Business Excellence, vol. 28, Issue 5-6, pp. 559-577, November 2015.

[7] P. K. Sari, A. Alamsyah, and S. Wibowo, "Measuring e-commerce service quality from online customer review using sentiment analysis", Journal of Physics: Conference Series, vol. 971, Issue 1, 2018.

[8] F. D. Orel and A. Kara, "Supermarket self-checkout service quality, customer satisfaction, and loyalty: Empirical evidence from an emerging market", Journal of Retailing and Consumer Services, vol. 21, Issue 2, pp, 118-129, March 2014.

[9] R. Kunimoto and R. Saga, "Causal Analysis of User's Game Software Evalution Using hLDA and SEM", IEEJ, vol. 135, Issue 6, pp. 602-610, 2015.

[10] K. Lee and C. You, "Assessment of airport service quality: A complementary approach to measure perceived service quality based on Google reviews", Journal of Air Transport Management, vol. 71, pp. 28-44, August 2018.

[11] L. Martin-Domingo, J. C. Martín, and G. Mandsberg, "Social media as a resource for sentiment analysis of Airport Service Quality(ASQ)", Journal of Air Transport Management, vol. 78, pp. 106-115, July 2019.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis", Journal of The American Society for Information Science, vol. 41, Issue 6, pp. 391-407, 1990.

[13] D. M. Blei, A. Y. Ng, J. B. Edu, and M. I. Jordan, "Latent dirichlet allocation", The Journal of Machine Learning Research, No. 3, pp. 993-1022, 2003.

[14] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process", Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 17-24, 2003.

[15] C. J. Hutto and E. Gilbert, "VADER: a parsimonious rule-based model for sentiment analysis of social media text", International AAAI Conference on Web and Social Media, 2014.

[16] L. Xiaozhou, Z. Zheying, and S. Kostas, "Sentiment-aware Analysis of Mobile Apps User Reviews Regarding Particular Updates", Proceedings of The Thirteenth Intetnational Conference on Software Engineering Advances, pp. 99-107, 2018.

[17] C. J. Anderson and W. D. Gerbing, "Strucutural equation modeling in practice: A review and recommended two-step approach.", Psychologcal Bulletin, vol. 103, No. 3, pp. 411-423, May 1988.

[18] A Kachites, "Mallet: A Machine Learning for Language Toolkit", http://mallet.cs.umass.edu, [retrieved: October, 2019].

[19] "Natural Language Toolkit – NLTK 3.4.4 document", https://www.nltk.org, [retrieved: October, 2019].

[20] R. Ihaka, R. C. Gentleman, "The R Project for Statistical Computing", https://www.r-project.org, [retrieved: October, 2019].

 ISBN: 978-1-61208-757-3

Identifying Obstacles in Data Sharing by Automatic Extraction of Problematic Points in Documents

Yan Wan
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: wanyan@bupt.edu.cn

Yalu Wang
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: 17888843379@163.com

Guanhao Chen
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: wangyalu@bupt.edu.cn

Jinping Gao
School of Economics and Management
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: doc.gao@126.com

***Abstract*— This paper aims to propose an automatic viewpoint extraction method using open data obstacle extraction as an example. Open data (data sharing) is very important because it reduces job repeatability and increase productivity and openness of work. However, open data in China is not as well developed as we wish. It is hindered by various problems, such as the willingness to share, the incompatible of data formats, etc. In order to identify different problems, then allocate to relevant parties to tackle these problems, we adopt an automatic extraction algorithm of natural language processing techniques, to automatically identify problematic points (obstacles) of data sharing from relevant literature. In this paper, we first construct a vocabulary for "obstacles", so that machines can find "obstacles" in literature more accurately. Then, an extraction algorithm combined with word2vec and Pointwise Mutual Information (PMI) is proposed, to automatically find the sentences that talk about "obstacles" of open data in documents. An experiment of this method is carried out and analyzed. It shows that the proposed method can be a very good tool for similar tasks that need to find viewpoint from a large amount of documents but cannot be done by simple keyword searches.**

Keywords-open data; feature extraction; data sharing obstacles.

I. INTRODUCTION

In recent years, with the rapid growth of data production and the extensive application of information technology, data volume is accumulated at an unprecedented speed. While data is constantly being produced at an unimaginable speed, it is also distributed in different institutions that own data, including governments, enterprises, scientific research institutions and other departments. How to make full use of this massive data is highly valued by government and researchers.

The main theme of big data era lies in data sharing. By analyzing and mining the value of data, we can save resources, improve the quality of product and make profits in ways that are more effective. Nowadays data openness is indispensable for achieving these purposes. The concept of data openness has been raised by Auer et al. [1], which is defined as "Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control." The shared data include government data, science data and enterprise data. In open data process, there exists various problems, which hinder data dissemination, data explore and value realization. In order to solve these problems, we need first to identify these problems then assign them to responsible parties to seek solutions. This paper focus on problem identification, by extracting problematic points from open data literature using natural language processing technology.

In most open data studies, scholars carried out their research by reading a large amount of documents. This is time-consuming and may miss some important points. This paper suggests an automatic sentence extraction method which finds out issues encountered in open data process automatically in a more comprehensive and efficient way. The method we propose is Problematic Points Extraction Model (PPEM). It combines word2vec model and Pointwise Mutual Information (PMI) based method. Its performance is validated by comparing it with word2vec and PMI.

In brief, this paper includes three parts:

a) Propose a sentence extracting model to identify the problematic points from documents discussing open data issues.

b) Evaluate the above model using text mining evaluation method.

c) Classify the found open data problematic points.

In Section II, a survey of the related work is described. In Section III, we introduced the method and the procedure we carried our research. In Section IV, the result of our proposed method is described and compared with other possible methods. Finally, in Section V, conclusion and future work are discussed.

II. RELATED WORKS

There exist many studies about the problems in open data process, but seldom do researchers focus on the automatic problems extraction. In this section, we examine three kinds of literature that contains open data, defect detection and vocabulary construction. The open data literature mainly demonstrates the research status

 ISBN: 978-1-61208-757-3

of open data problems. Then we review the study of defect detection, which is a reasonable reference to our research. As we employ the vocabulary construction method to extract problematic points, we also make a brief introduction of this technology.

A. Open Data

In today's society, customers using mobile phone to accomplish many activities in their everyday life. Meanwhile, their personal information and behavior data has been stored by the service provider and related institute such as government departments, banks, enterprises and so on. The massive data stored in the above places has significant value for the development of economics and enhancement of people's living standard. However, the reality is that most of the data remains unused and the value of data is being wasted. It is time that we undertake some strategies for data liberalization.

In the process of data liberalization, the government as the master of a large number of high-value data resources and the standardization of the use of information resources are the main force to promote the open sharing of data. To promote the e-government programs, Dawes designed an electronic government information access programs in his research [2]. In Conradie and Choenni's research [3], they regarded the way of data acquisition, storage and use, and the suitability of data openness, as crucial indicators for open data releasing. They conducted interviews and workshops and finally classified the barriers to data release as fear of false conclusions, financial effects, opaque ownership and unknown data locations, low priority [3].

B. Difficulty Discovery

In this paper, we define the concept of problematic words, which are problem-oriented words in sentences to indicate problematic sentences. The research conducted by Abrahams et al. [4] in Virginia Tech provides a good example for us. In their research, they first applied automatic defect discovery approach on discussion threads of vehicle forums and proposed Vehicle Defect Discovery System [4]. They also defined 'smoke words' concept to better recognize vehicle defects. Subsequently, they undertook another study about automotive defect and consumer electronics defect, and proposed Social Media Analytic framework using Text(SMART) for Quality Management (QM) tasks [5]. Later in 2017, they extended their methods into the defect discovery of toys and dishwasher appliances [6][7].

C. Vocabulary Construction

Vocabulary construction mostly appears in sentiment analysis tasks, by doing so, researchers can distinguish the linguistic unit between positive and negative polarities. Vocabulary, also called lexicon, can be divided into two categories, domain-oriented and domain-independent. Kanayama and Nasukawa [8] proposed an unsupervised lexicon building method, in which they used the "polar atoms" as the linguistic unit and calculate context coherency. Our research aims at the problems exist in open data domain, and we suppose to establish a domain-oriented lexicon.

III. METHODS

Using keywords "open data", and then eliminating duplicate and too short articles, 486 documents are obtained from CNKI database (the most popular Chinese academic database). 436 documents are used to train the Word2vec model and the remaining 50 documents are for experiment. In this section, the PPEM's procedures, including data pre-processing, lexicon construction and problematic points extraction, will be explained in detail.

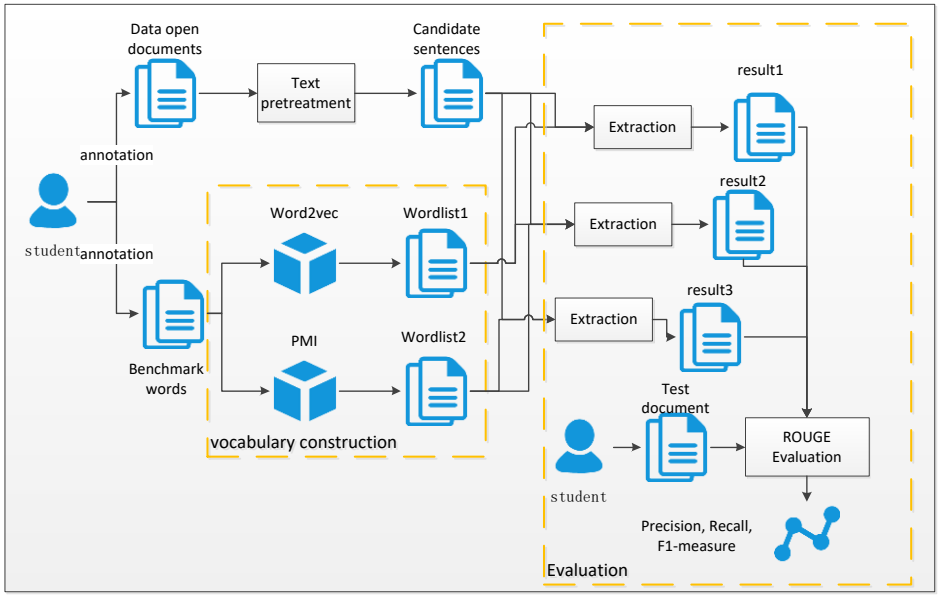

Figure 1. Procedure of problematic points extraction model (PPEM)

TABLE I. PROBLEMATIC WORDLIST CONSTRUCTION

Benchmark words	Problematic words
问题(problem)	挑战(challenge),难题(difficulty), 障碍(obstacle),后顾之忧(future trouble)
缺乏(shortage)	缺少(shortage), 不够(not enough), 薄弱(weak), 缺失(lack)
不足(shortcoming)	诸多(numerous)，障碍(obstacle)，明显(obvious)，差异(difference), 突出(prominent)
困难(difficulty)	现象(phenomenon), 严重(serious), 安全隐患(potential safety hazard), 难(difficult)
差距(gap)	差异(difference), 差别(difference), 明显(obvious),成就(achievement), 落后（fall behind）
无法(cannot)	不能(unable), 难以(difficult to), 能够(able to), 很难(hard)
尚未(not yet)	并未(wasn't),尚(yet), 尽管(despite), 较晚(later), 虽然(although)
阻碍(hinder)	制约(restrict), 起到(serve as), 忽视(ignore), 严重(serious), 改变(alter)
挑战(challenge)	机遇(opportunity), 难题(puzzle), 问题(problem), 面临(confront)

A. Text Pre-processing

After the collection of open data documents, the first step is text pre-processing to clean and formalize the texts. First, we eliminate the graphs, tables and annotations of the text. Next, we write a computer program to remove the literature serial numbers, non-Chinese characters, and other redundant content that influence the quality of analysis. For stop word elimination and word segmentation, we employ the python modules, *jieba* and *snownlp*, which are useful natural language processing tools.

B. Lexicon Construction and Problematic Point Extraction

As we can see in Figure 1, next to the text pre-processing is vocabulary construction step. In this step, we introduce Word2vec and PMI to calculate the similarity of words. Traditionally, for information extraction, many researchers use some kind of algorithm to score the sentences and select the sentences that are highly ranked. In our study, we suppose to collect the similarity scores and select the top n words to form the problematic vocabulary. The wordlist constructed by word2vec and PMI is named as wordlist1 and wordlist2, illustrated in table 3. In the next, we can use these scores to calculate the sentence score, which contain words in above vocabulary. The sentence scores generate by different models are certainly different. In Figure 1, we define the outcome of PPEM, Word2vec and PMI as result1, result2 and result3.

The PPEM model is organized by a simple combining algorithm and the wordlists generated by word2vec and PMI. The metric of PPEM is calculated as in (1).

$$Weight(w) = \rho\ Weight_{w2v}(w) + (1-\rho) Weight_{PMI}(w) \quad (1)$$

In (1), ρ weighs the significance of word2vec, and 1-ρ weighs the significance of PMI. We can adapt optimal parameters by repeated experiment to our extraction model.

C. ROUGE Evaluation

The ROUGE evaluation method is extensively used in text mining field, which represents Recall-Oriented Understudy for Gisting Evaluation. This evaluation model operates through comparing candidate result with reference result [8][9]. The candidate result is generated by computer, while reference result is generated by experts. To validate the capability of PPEM model in a relatively convincing way, we choose three kind of metrics in ROUGE evaluation system, namely ROUGE-1, ROUGE-2, ROUGE-L. ROUGE-1 and ROUGE-2 represents the overlap of unigrams and bigrams between candidate result and reference result. ROUGE-L metric represents the overlapping longest common subsequence.

IV. RESULT AND EVALUATION

An experiment is designed to extract the problematic points by sentence extracting model PPEM, which employs the word2vec model and PMI indices. We use python program to realize the methods of word2vec and PMI, and ask three students who major in management science and engineering to accomplish the manual annotation task.

TABLE II. ROUGE METRICS OF PPEM, WORD2VEC, PMI

Model		Word2vec	PMI	PPEM
ROUGE-1	**P**	58.6	60.3	61.6
	R	82.8	80	87
	F1	68.6	68.8	72.1
ROUGE-2	**P**	39.3	42.0	44.0
	R	55.5	55.6	62.1
	F1	46.2	47.9	51.5
ROUGE-3	**P**	51.4	53.5	55.0
	R	72.6	71.0	77.7
	F1	50.2	51.0	54.4

As is shown in Table 2, we calculate the precision, recall, F1-measure of three metrics, which represented by P, R, F1. Precision is the percentage of true positive samples comparing with the true samples. In contrast, recall is the percentage that true positive samples divided by all positive samples. F1-measure represents the harmonic mean of precision and recall.

The output data shows that the sentence extracted by word2vec has higher recall but is lower in precision than PMI. Comparing the F1-measure, it can be summarized that the PPEM model have better performance in all these metrics, which denote that

 ISBN: 978-1-61208-757-3

PPEM is a more suitable approach for problematic points extraction.

V. CONCLUSION AND FUTURE WORK

In this paper, we have studied the problems in open data field by employing word2vec and PMI techniques. The main contributions of our research are as follows:

(1) applying natural language processing approach to automatically extract the problems in open data field from a large number of documents.
(2) proposing a new model PPEM which combines word2vec and PMI. Evaluation of the performance of the model is conducted.

A vocabulary for problem discovery is constructed to improve the performance of PPEM.

By observing the experiment results, we can draw conclusions as follows:

The PPEM model performs better than extracting by word2vec or PMI alone. In other words, it is a reasonable choice for researchers to apply PPEM model to coping with problems in specific domains. The output sentences extracted by PPEM represent the main problems in open data field in China. To summarize, the main problematic points of open data focus on four aspects, as shown in Table 3.

TABLE III. PROBLEMS IN DIFFERENT FIELDS

Field	Problems
Data source	• government neglect of data liberalization • weak data storage • lack of personnel in charge of building open data departments • scattered data of enterprises • lack of a unified open standard
Data dissemination	• privacy protection • inconsistent standard • imperfect data legislation
Data analysis	• short of data analysis professionals • the task of data analysis is not yet clear
Data application	• big data technology has not been popularized among the general public • less data open enterprises • enterprises lack of funds

In general, in order to promote further open data, issues should be considered from national strategic level and a special leading group can be set up to coordinate data openness. Local governments need to implement the opening-up policy, strengthen the construction of an open platform for data and eliminate the isolation of information. Legislative and judiciaries need to promote data-related legislation and regulate the way data is used. Large-scale enterprises need to regulate data formats and desensitize sensitive data. They should abide by the data usage rules, not use the data for illegal activities, and on the other hand, enhance their understanding of the value of data and their ability to obtain and analyze data. Therefore, the process of opening up the data requires the joint promotion of the three parties, including the government, enterprises and individuals, in order to continuously move forward. However, because this paper is mainly concerning the data analysis methodology, the open data issue itself has not been explored enough and shall be discussed further in future work.

ACKNOWLEDGMENT

This research partly supported by Natural Science Foundation of China (Nos. 71874018, 71471019, 71772017).

REFERENCES

[1] S. Auer, et al., "DBpedia: A Nucleus for a Web of Open Data," The Semantic Web: the sixth International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC + ASWC, Nov. 2007, pp. 722-735, doi: 10.1007/978-3-540-76298-052.

[2] S. S. Dawes, T. A. Pardo and A. M. Cresswell, “Designing electronic government information access programs: a holistic approach,” Government Information Quarterly, vol. 21, 2004, pp. 3-23, doi:10.1016/j.giq.2003.11.001.

[3] P. Conradie and S. Choenni, “Exploring process barriers to release public sector information in local government,” The twelfth International Conference on Theory and Practice of Electronic Governance Icegov(ICEGOV), ACM, Oct. 2012, pp. 5-13, doi:10.1145/2463728.2463731.

[4] A. S. Abrahams, J. Jiao, G. A. Wang and W. G. Fan, “Vehicle defect discovery from social media,” Decision Support Systems, vol. 54, Dec. 2012, pp. 87-97, doi: 10.1016/j.dss.2012.04.005.

[5] A. S. Abrahams, W. Fan, G. A. Wang and W. G. Fan, “An integrated text analytic framework for product defect discovery,” Production & Operations Management, vol. 24, Sept. 2014, pp. 975-990, doi: 10.1111/poms.12303.

[6] M. Winkler, A. S. Abrahams, R. Gruss and J. P. Ehsanib, “Toy safety surveillance from online reviews," Decision Support Systems, vol. 90, Oct. 2016, pp. 23-32, doi:10.1016/j.dss.2016.06.016.

[7] D. Law, R. Gruss and A. S. Abrahams, “Automated defect discovery for dishwasher appliances from online consumer reviews,” Expert Systems with Applications, vol. 67, Jan. 2017, pp. 84-94, doi:10.1016/j.eswa.2016.08.069.

[8] H. Kanayama and T. Nasukawa, “Fully Automatic Lexicon Expansion for DomainOriented Sentiment Analysis,” Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, ACL, Jan. 2006, pp. 355-363, doi:10.3115/1610075.1610125.

[9] C. Flick, “ ROUGE: A Package for Automatic Evaluation of summaries," Proceedings of the Workshop on Text Summarization Branches Out(WAS), Jan. 2004, pp. 10, doi:doi:http://dx.doi.org/.

 ISBN: 978-1-61208-757-3

Identifying Latent Toxic Features on YouTube Using Non-negative Matrix Factorization

Adewale Obadimu
Department of Computer Science
University of Arkansas at Little Rock
Little Rock, USA
Email: amobadimu@ualr.edu

Esther Mead
Department of Information Science
University of Arkansas at Little Rock
Little Rock, USA
Email: elmead@ualr.edu

Nitin Agarwal
Department of Information Science
University of Arkansas at Little Rock
Little Rock, USA
Email: nxagarwal@ualr.edu

***Abstract*—Toxic behavior, in its various forms, often disrupts constructive discussions in online communities. The proliferation of smart devices and mobile applications has further exacerbated these nefarious acts on various social media platforms. Largely, toxic behavior is regulated by human moderators employed by the platform operators. However, given the volume and speed of content posted on online platforms, identifying and deterring these behaviors remains challenging. In this study, we propose a Non-negative Matrix Factorization (NMF) technique for predicting commenter toxicity on YouTube. We utilized the YouTube Data API to collect data from the Cable News Network (CNN) channel on YouTube. Our final dataset consists of 144 videos, 243,344 commenters, and 421,924 comments. We then utilized Google's Perspective API to assign a toxicity score to each comment. We used the resultant dataset to create a commenter toxicity score prediction model. We tested our proposed NMF model against other popular prediction methods, comparing speed of model execution and the common Root-Mean-Square-Error (RMSE) accuracy metric. This work sets the stage for a richer, more detailed analysis of toxicity on various online social media networks.**

Keywords—Toxicity; Tonality Analysis; YouTube; Social Media; Language Model.

I. INTRODUCTION

The advent of computer-mediated forms of interactions has posed several conceptual and practical challenges [1]. With emergent norms and conventions, the Web, more than any other medium, has offered lowered communication thresholds and a broadened geographical scope of human interactions [2]. However, despite the myriad advantages of utilizing this medium to connect to like-minded individuals, a consensus is emerging suggesting the presence of malicious actors, otherwise known as trolls [3]. These actors (hereafter referred to as *toxic users*) thrive on disrupting the norms of a given platform and causing emotional trauma to other users [4]. In this study, similar to extant literature, we give an operational definition of toxicity as "the usage of rude, disrespectful, or unreasonable language that will likely provoke or make another user leave a discussion" [3][5][6][7]. Therefore, in this regard, toxicity analysis is different from sentiment analysis, which is the attempt to assign sentiment scores of positive, neutral, and negative to text data.

Social media was once perceived as a liberating platform but is now riddled with various forms of toxicity [5]. A report by the Pew Research Center indicated that 73% of adult Internet users have seen someone harassed online, and 40% have experienced it personally [3][8]. Another survey by Duggan [8] highlighted that 19% of teens reported that someone has written or posted malicious or embarrassing things about them on social networking sites. Due to the growing concerns about the impact of online harassment, many platforms are taking several steps to curb this phenomenon [5][6][9][10]. For instance, on YouTube, a user can simply activate the safety mode to filter out offensive language [8]. Wikipedia has a policy of "Do not make personal attacks anywhere in Wikipedia" [5]. Likewise, platforms like CNN.com have moderators that reportedly remove over one in five comments that violate community guidelines on any given day [3]. The aforementioned are a few examples that highlight the negative impacts of toxic behavior on the community. Toxic behavior, if not curbed at the initial stage, can have a ripple effect. It can dissuade other people from joining a community by perceiving the community as a hostile environment [5].

According to Alexa, the web traffic monitoring service owned by Amazon, YouTube is the second most popular website globally with over 300 hours of videos uploaded every minute and 5 billion videos watched every single day [10]. While several studies have attested to the widespread manifestation of toxicity within comments on YouTube [5][11][12], the burden of access to such a large dataset has made a permanent extraction of toxic users challenging. Due to the extensive exploitation of these platforms by toxic users, automatic detection and extraction of toxicity has become a pressing need [9].

Despite the rich vein of academic research on identifying various forms of online toxicity, tackling these behaviors at scale remains surprisingly challenging [3]. This, along with the immensity of the amount of data and the speed with which the data is generated and shared, motivated us to propose a Non-negative Matrix Factorization (NMF) technique for unraveling latent toxic commenter features on YouTube. The intuition behind

 ISBN: 978-1-61208-757-3

using matrix factorization to solve this problem is that there should be some latent features that determine the toxicity of a commenter on a given video. For instance, two commenters may have higher toxicity on a video if they both dislike who or what the video is talking about, or perhaps, if they both mutually dislike the genre of the video. Hence, if we can automatically discover these latent features, we should be able to predict the toxicity of a certain commenter on a certain video. This work is novel in that the NMF approach has not previously been applied to this type of problem. The contribution of this work is to help understand the relationship between the impact of toxicity of a video on the comments, to enable us to predict the likely toxicity of a commenter based on their past history, and to allow us to determine what kind of comments a video will generate based on prior toxicity matrix. We selected the CNN news channel's "*must-see moments*" videos section as an experimental dataset because it is rich in various forms of behaviors.

The remainder of this paper is set out as follows. In Section 2, we present a theoretical formulation of the problem. Then, we give a brief review of extant literature that are most germane to our discussion in Section 3. Next, our methodology is described in Section 4, and the findings are discussed. Section 5 provides conclusions including the limitations of our work, and ideas for future work.

II. PROBLEM FORMULATION

We pose NMF as an optimization problem where, in addition to minimizing the reconstruction error of the commenter-video toxicity matrix, we also require that the factors capture prior knowledge as much as possible. The intuition behind using this approach is that there should be some latent features (characteristics that are not directly observed [30]) that determine the toxicity of a given commenter on a specific video. In trying to discover these latent features, we assume that the number of features would be smaller than the number of commenters and the number of videos. To validate the efficacy of this approach, we applied this method to a real-world YouTube dataset, making our work the first to conduct toxicity analyses using NMF on YouTube.

Given a set of commenters $C = \{c_1, \dots, c_N\}$ and a set of videos $V = \{v_1, \dots, v_M\}$, the toxicity expressed by these commenters on all the videos can be expressed in a toxicity matrix $T = [T_{c,v}]_{N \times M}$. In this matrix, $T_{c,v}$ represents the average toxicity of a commenter c on a video v and it is bounded in the range of $[0,1]$. Our objective in this study is as follows: Given a commenter $c \in C$ and a video $v \in V$ for which $T_{c,v}$ is unknown, predict the toxicity for c on video v using T. T is asymmetric and usually very sparse.

Let $P \in R^{K \times N}$ and $Q \in R^{K \times M}$ be latent commenter and video feature matrices, with column vectors P_c and Q_v representing K-dimensional commenter-specific and video-specific latent feature vectors of commenter c and video v, respectively. The resulting dot product ${P_c}^T Q_v$ captures the interaction between commenter c and video v. This product approximates commenter c's toxicity on video v, and it is denoted by $\hat{T}_{c,v}$ as shown in (1).

$$\hat{T}_{c,v} = {P_c}^T Q_v \quad (1)$$

To learn the latent feature vectors (P_c and Q_v), we minimize the regularized squared error (2) on the set of known toxicity using stochastic gradient descent.

$$min \sum_{(c,v)\in\Im} (T_{c,v} - \hat{T}_{c,v})^2 + \lambda(||(P_c|| + ||Q_v||) \quad (2)$$

Here, $\Im$ is the set of the (c, v) pairs for which $T_{c,v}$ is known (the training set). The conditional probability of the observed toxicity (3) is defined as:

$$p(T|P, Q, \sigma_T^2) = \prod_{c=1}^{N} \prod_{v=1}^{M} \left[N(T_{c,v}|({P_c}^T Q_v), \sigma_T^2\right] \quad (3)$$

where $N(x|\mu, \sigma_T^2)$ is a normal distribution with mean μ and variance σ^2.

III. RELATED WORK

This section discusses the two categories of related work that correspond to our research. The first category includes works that have attempted to identify toxicity on social media. The second category includes the works that have attempted to utilize matrix factorization techniques.

A. Identifying toxicity on Social Media

Threads of extant literature on antisocial behavior suggest that toxicity, in its various forms, oftentimes disrupts constructive discussions in an online community [3][5][12][13]. Several researchers have tried to identify and suggest ways to mitigate toxicity in a community [5][13]. Using data collected via crowdsourcing, Wulczyn et al. [5] employed machine learning techniques, such as linear regression and multilayer perceptron to analyze personal attacks on social media. A study by Martens et al. [6] utilized Natural Language Processing techniques to detect the emergence of undesired and unintended behavior in online multiplayer games. Research by Chen et al. [14] and Yin et al. [15] used a set of regular expressions, n-grams, and supervised learning techniques to detect abusive language. Sood et al. [16] combined lexical and parser features to identify offensive language

 ISBN: 978-1-61208-757-3

in comments extracted from a social news site. Davidson et al. [17] presented a dataset with three kinds of comments: hate speech, offensive but non-hateful speech, and neither. Hosseini et al. [18] demonstrated the vulnerability of most state-of-the-art toxicity detection to adversarial inputs. Despite the rich vein of academic work on toxicity detection, there is a need for systematic research that focuses on detecting, identifying, and categorizing toxicity at scale.

B. Matrix Factorization Techniques

Our work using NMF was heavily inspired by the need for an approach that scales to the huge amount of streaming data on YouTube. The idea behind matrix factorization is to decompose an interaction matrix into a product of two lower dimensionality rectangular matrices while minimizing the error associated with the decomposition [19]. This technique has been used extensively in recommendation systems to discover latent features underlying the interactions between users and items ratings [20][21]. Yang et al. [22] applied a matrix factorization technique for developing a model-based community detection algorithm that detects densely overlapping communities in a network. Ma et al. [20] utilized a probabilistic matrix factorization technique to solve the data sparsity and poor prediction accuracy problems by employing both users' social network information and rating records to perform recommendation. Zhao et al. [23] employed a matrix factorization method on each of the communities in a unidirectional social network. By advancing previous work, Jamali et al. [24] proposed a matrix factorization technique with trust propagation for recommendation systems. Chen et al. [25] proposed a novel social recommendation method that fuses user's social status with homophily using a matrix factorization technique. Peng et al. [21] proposed a social trust and segmentation-based matrix factorization recommendation algorithm. Ozer et al. [26] leveraged matrix factorization techniques to uncover political networks on Twitter. However, to the best of our knowledge, our work is the first to apply matrix factorization to unravel toxic features on YouTube.

IV. METHODOLOGY

Our methodology (Fig. 1) consists of three phases: 1) data collection and data processing; 2) data preparation and toxicity assignment; and, 3) matrix factorization of commenter-video toxicity matrix.

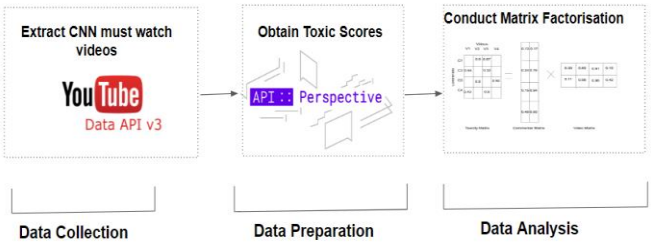

Figure 1. Research methodology.

A. Data Collection

To build our dataset for illustrating our proposed method, we first utilized Google's YouTube API to extract videos' and commenters' data from the "*must see moments"* video playlist from the CNN channel. To reduce "noise" in extracted data, several data processing steps were subsequently performed including data formatting, data standardization and data normalization using the Python programming language. Our final dataset consists of 144 videos, 243,344 commenters, and 421,924 comments.

B. Data Preparation

The next step in our methodology was to assign toxicity scores to each comment in the dataset. To accomplish this, we leveraged a classification tool called Perspective API [27], which was developed by Google's Project Jigsaw' and 'Counter Abuse Technology' teams (Table I). This model uses a Convolutional Neural Network (CNN) trained with word-vector inputs to determine whether a comment could be perceived as "toxic" to a discussion [27][28]. The API returns a probability score between 0 and 1, with higher values indicating a greater likelihood of the toxicity label. Table I is an excerpt of the resultant toxicity dataset where toxicity scores have been assigned to each comment.

TABLE I. CONVENIENCE SAMPLING OF FIVE (5) TOXIC COMMENTS IN OUR DATASET.

S/N	*Comment*	*Overall Toxicity*
1	SHUT UP YOU OLD FOOL !	0.97
2	dumba** liberal kids..#cnnsucks	0.74
3	JIM your a special kind of STUPID..	0.97
4	Brian, you are so dumb.	0.96
5	Stupid dumb gun lover.	0.78

C. Data Analysis

Before applying algorithms for toxicity prediction, we conducted some preliminary data analysis on our YouTube CNN video comments dataset. Fig. 2 shows the distribution of overall toxicity in our dataset. Although most of the comments (80%) were assigned a toxicity score of less than 0.5, a significant portion (20%) of the

 ISBN: 978-1-61208-757-3

comments were assigned a toxicity score of greater than 0.5 (over 84,000 comments). The mean toxicity score was about 0.27.

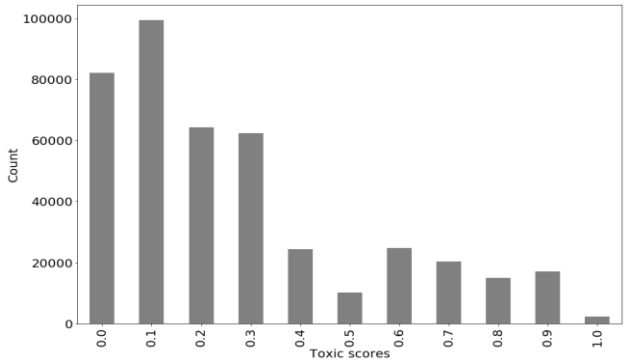

Figure 2. Distribution of overall toxicity.

Most of the toxic comments within the CNN dataset were categorized as being some type of “INSULT”—a negative comment towards a person or a group of people, followed closely by “THREAT”—an intention to inflict pain, injury, or violence against an individual or group (Fig. 3).

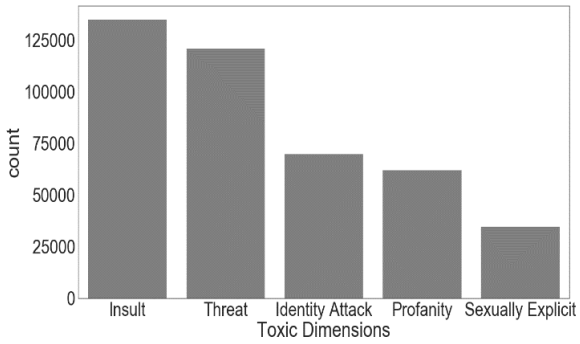

Figure 3. Distribution of toxicity by toxicity dimension.

Fig. 4 shows the box plots for each of the toxicity dimensions and reveals that the “INSULT” and “PROFANITY”—the usage of swear/curse words, or other obscene/profane language—had the widest spread with regard to the distribution of toxicity scores. However, compared to other toxicity dimensions, sexually explicit comments were the least severe.

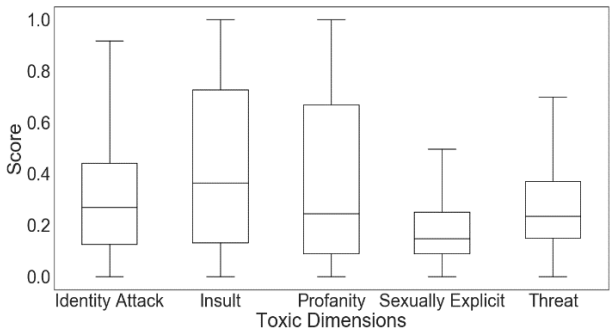

Figure 4. Box plot for each toxicity dimension.

Word clouds generated from the dataset reveal some of the content that relates to these five toxicity dimensions (Fig. 5-7). As can be seen in the content of each word cloud, some of the categories are more obvious than others; for example, the prevalence of terms such as, “kill” and “shooting” within the “THREAT” toxicity dimension (Fig. 6). Similarly, the prevalence of terms such as, "black", "white", "fake", "Democrat", and "Republican" within the "IDENTITY ATTACK" toxicity dimension (Fig. 7) give indications of why these comments were categorized as belonging to this specific toxicity dimension. Some of the word clouds that we generated could not be included in this work because of the obscenity, profanity, and sexually explicit terms that they revealed. These word clouds corresponded to the toxicity dimensions of "PROFANITY" and "SEXUALY EXPLICIT".

Figure 5. “INSULT” word cloud.

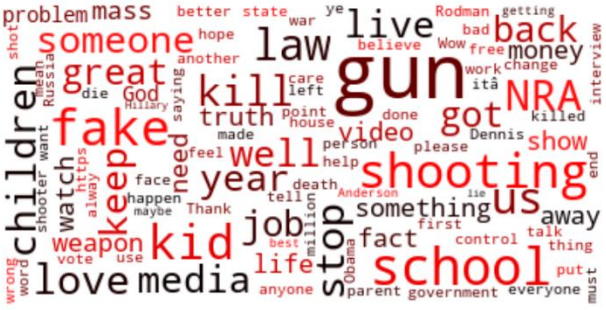

Figure 6. “THREAT” word cloud

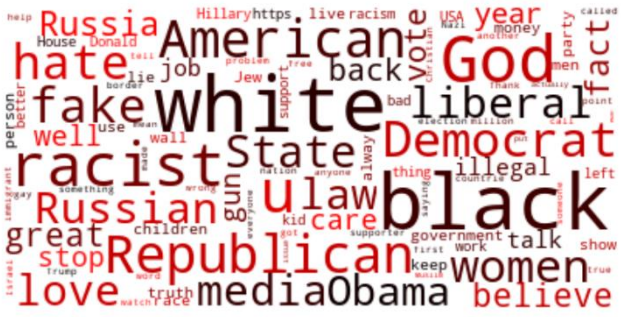

Figure 7. “IDENTITY ATTACK” word cloud

D. Evaluation of the approach

The next step in our methodology was to apply the algorithms for toxicity prediction and compare the results. The most commonly used metrics for assessing the effectiveness of predictive methods, such as the proposed NMF algorithm are the mean squared error and Root-Mean-Square-Error (RMSE), the latter having been used in the Netflix Prize [19]. RMSE is a measure that can be used to compare predictions against real data. The smaller the RMSE value, the better the model. The error was computed, and gradient descent was performed to minimize the error. We used a random 70/30 split of our dataset to create training and test sets and applied 5-fold cross-validation. Using *Surprise Python Scikit Package* [29], we compared the result of the NMF approach with other techniques: CoClustering (Collaborative filtering algorithm) and NormalPredictor (Algorithm based on the normal distribution of the training set) (Table II).

TABLE II. EVALUATING RMSE ON 5 SPLIT(S). SMALLER THE VALUE THE BETTER THE MODEL.

Algorithm	*Mean RSME*
NMF **Improved %**	**0.28** **17.85%**
CoClustering	0.33
NormalPredictor	0.34

Compared to CoClustering and NormalPredictor, the NMF approach performs better in terms of accuracy (having the lowest RMSE). Table III shows that the NMF approach outperformed NormalPredictor based on computation time, but not CoClustering. The mean computation time for NMF was 0.54 seconds, while NormalPredictor took 0.56 seconds and CoClustering took 0.45 seconds. This experiment was conducted on a machine with Intel(R) Xeon(R) CPU E7-8893 v4 @ 3.20GHz 3.19 GHz (4 processors) and 3.25 TB RAM. The complexity analysis indicates that our approach can be applied to very large datasets since it scales linearly with the number of observations.

V. CONCLUSION

In this study, we addressed the problem of identifying and predicting toxicity on online social media networks. We chose to focus on user comments posted on a sample of CNN videos posted on YouTube as a case study for illustrating our methodology. The challenges in addressing this problem include an ongoing issue of balancing freedom of expression with curtailing harmful content. The contribution of this paper is that it outlines a scalable methodology for first identifying toxicity within commenter text data posted on an online social media network and then predicting the toxicity levels of each commenter. We found that the proposed NMF-based approach performed better than some other techniques for predicting toxicity scores in terms of and accuracy and has the potential to perform better in terms of computation time. These findings demonstrate how the presence of toxicity in a set of text corpora can be identified, categorized, and measured, and how those metrics can be used for prediction. Our findings advance the research in this topic in that this analysis serves as a steppingstone in a long line of future work that can be done to better understand the origin, propagation and impact of toxicity on online social media networks. The general implications of this work are that by systematically assessing toxicity, this work sets the stage for developing a richer, more robust model for understanding the flow of toxicity on various online social media networks and for developing tools for toxicity control and prevention. There are a few limitations to this work, however. One is that it is challenging to model a cold-start commenter—a commenter that has never posted a comment—as new users will not have an initial toxicity score. Additionally, the commenter-video toxicity matrix can become very sparse with increasing volume of data thereby increasing the reconstruction error. Our immediate next steps will investigate possible solutions to address the limitations. Long term future work includes running experiments on multiple datasets and considering other variables to determine whether this can improve prediction accuracy. We anticipate applying the NMF method to data from other online social media networks. For instance, a sophisticated version of the NMF approach can be used to capture the bidirectional latent relations between user's toxicity preferences across domain through transfer learning. Future work should also attempt to model the spread of toxicity on online social media networks and determine whether these metrics and predictions can be used for developing preventive measures.

ACKNOWLEDGMENT

This research is funded in part by the U.S. Office of Naval Research (N00014 - 10 - 1 - 0091, N00014 - 14 - 1 - 0489, N00014-15-P-1187, N00014-16-1-2016, N00014 - 16 - 1- 2412, N00014-17-1-2605, N00014-17-1-2675, N00014-19-1-2336), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, and the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily

 ISBN: 978-1-61208-757-3

reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

REFERENCES

[1] M. Warschauer, "Computer-mediated collaborative learning: Theory and practice." The Modern Language Journal, vol. 81, no. 4, pp. 470-481, 1997.

[2] E. Eleanor, "Hyperbole over cyberspace: Self-presentation and social boundaries in Internet home pages and discourse," The Information Society, vol. 13, no. 4, pp. 297-327, 1997.

[3] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions," Proc. ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW17), ACM Press, Feb. 2017, pp. 1217–1230, doi: 10.1145/2998181.2998213.

[4] S. Lee and H. Kim, "Why people post benevolent and malicious comments online," Communications of the ACM, vol. 58, no. 11, pp. 74–79, 2015.

[5] E. Wulczyn, T. Nithum, and L. Dixon, "Ex machina: Personal attacks seen at scale," Proc. 26th International Conference on World Wide Web (WWW 2017) ACM, [month] 2017, pp. 1391-1399, ISBN: 978-1-4503-4914-7, doi: 10.1145/3038912.3052591.

[6] M. Martens, S. Shen, A. Iosup, and F. Kuipers, "Toxicity detection in multiplayer online games," International Workshop on Network and Systems Support for Games (NetGames 2015) IEEE/ACM, De. 2015, pp. 1–6, ISBN: 978-1-5090-0068-5.

[7] A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal, "Identifying Toxicity Within YouTube Video Comment," Proc. International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2019) Springer, Jul. 2019, pp. 214-223, ISSN: 0302-9743, ISBN: 978-3-030-21740-2, doi: 10.1007/978-3-030-21741-9.

[8] M. Duggan, "Online Harassment", Pew Research Center. [Online]. Retrieved: August 6, 2019. Available from: http://www.pewinternet.org/2014/10/22/online-harassment/

[9] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," Proc. International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing (PASSAT-SOCIALCOM 2012), IEEE/ASE, Sept. 2012, pp. 71–80, ISBN: 9781467356381 and 978-0-7695-4848-7.

[10] M. N. Hussain, S. Tokdemir, S. Al-khateeb, K.K. Bandeli, and N. Agarwal, "Understanding digital ethnography: socio-computational analysis of trending YouTube videos," Proc. International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2018) Springer, Jul. 2018, Late Breaking Paper. [Online]. Retrieved: August 7, 2019. Available from: http://sbp-brims.org/2018/proceedings/papers/latebreaking_papers/LB_14.pdf

[11] M. N. Hussain, S. Tokdemir, N. Agarwal, and S. Al-Khateeb, "Analyzing disinformation and crowd manipulation tactics on YouTube." Proc. International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2018), IEEE/ACM, Aug. 2018, pp. 1092-1095, doi: 10.1109/ASONAM.2018.8508766.

[12] C. Nobata, J Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," Proc. 25th International Conference on World Wide Web (WWW 2016), ACM, Apr. 2016, pp. 145-153, ISBN: 978-1-4503-4143-1.

[13] Q. Cao, X. Yang, J. Yu, and C. Palow, "Uncovering large groups of active malicious accounts in online social networks," Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS 2014), ACM, Nov. 2014, pp. 477-488, ISBN: 978-1-4503-2957-6.

[14] Y. Chen, Y. Zhou, and S. Zhu, H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," Proc. International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing (PASSAT-SOCIALCOM 2012), IEEE/ASE, Sept. 2012, pp. 71–80, ISBN: 978-0-7695-4848-7, doi: 10.1109/SocialCom-PASSAT.2012.55.

[15] D. Yin et al., "Detection of harassment on web 2.0," Proc. Content Analysis in Web 2.0 (CAW2.0 2009), EPrints, Apr. 2009, pp. 1-7. [Online]. Retrieved: August 7, 2019. Available from: http://www2009.eprints.org/255/6/Yin_etal_CAW2009.pdf

[16] S. O Sood, J. Antin, and E. Churchill, "Using crowdsourcing to improve profanity detection," Proc. Spring Symposium Series of the Association for the Advancement of Artificial Intelligence (AAAI 2012), AAAI Press, Mar. 2012, ISBN 978-1-57735-555-7. [Online]. Retrieved: August 7, 2019. Available from: https://www.aaai.org/ocs/index.php/SSS/SSS12/paper/view/4256/4698

[17] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Proc. Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017), AAAI Press, May 2017, ISBN 978-1-57735-788-9. [Online]. Retrieved: August 7, 2019. Available from: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665/14843

[18] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving Google's Perspective API Built for Detecting Toxic Comments," arXiv preprint, Feb. 2017, arXiv:1702.08138. [Online]. Retrieved: August 7, 2019. Available from: https://arxiv.org/pdf/1702.08138.pdf

[19] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30-37, 2009.

[20] H. Ma, H. Yang, M. Lyu, and I. King, "Sorec: social recommendation using probabilistic matrix factorization," Proc. 17th ACM conference on Information and knowledge management (CIKM 2008), ACM, Oct. 2008, pp. 931-940, ISBN: 978-1-59593-991-3, doi: 10.1145/1458082.1458205.

[21] W. Peng and B. Xin, "SPMF: A Social Trust and Preference Segmentation-based Matrix Factorization Recommendation Algorithm," arXiv preprint, Mar. 2019, arXiv:1903.04489. [Online]. Retrieved: August 7, 2019. Available from: https://arxiv.org/pdf/1903.04489.pdf

[22] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," Proc. Sixth ACM International Conference on Web Search and Data Mining, ACM, Feb. 2013, pp. 587-596, ISBN: 978-1-4503-1869-3, doi: 0.1145/2433396.2433471.

[23] G. Zhao, M. Lee, W. Hsu, W. Chen, and H. Hu, "Community-based user recommendation in uni-directional social networks," Proc. 22nd ACM International Conference on Information & Knowledge Management, ACM, Oct. 2013, pp. 189-198, ISBN: 978-1-4503-2263-8, 10.1145/2505515.2505533.

[24] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks." Proc. Fourth ACM Conference on Recommender Systems, ACM, Sep. 2010, pp. 135-142, ISBN: 978-1-60558-906-0, doi: 10.1145/1864708.1864736.

[25] R. Chen, et al., "A Novel Social Recommendation Method Fusing User's Social Status and Homophily Based on Matrix Factorization Techniques," IEEE Access, vol. 7, pp. 18783-18798, 2019, doi: 10.1109/ACCESS.2019.2893024.

[26] M. Ozer, N. Kim, and H. Davulcu, "Community detection in political Twitter networks using Nonnegative Matrix Factorization methods," Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and

 ISBN: 978-1-61208-757-3

Mining (ASONAM 2016), ACM, Aug. 2016, pp. 81-88, ISBN: 978-1-5090-2846-7.

[27] Perspective. [Online]. Retrieved: August 6, 2019. Available from: http://perspectiveapi.com/#/

[28] S. Bay et al., "Responding to Cognitive Security Challenges," NATO STRATCOM COE. [Online]. Retrieved: August 6, 2019. Available from:| https://www.stratcomcoe.org/responding-cognitive-security-challenges.

[29] N. Hug. "Surprise: A Python scikit for recommender systems." [Online]. Retrieved: August 6, 2019. Available from: http://surpriselib.com/

[30] M. Graus, M. Willemsen, and L. Meesters, "Understanding the latent features of matrix factorization algorithms in movie recommender systems," Eindhoven University of Technology, 2011. [Online]. Retrieved: August 7, 2019. Available from: https://pure.tue.nl/ws/files/47007020/712626-1.pdf

 ISBN: 978-1-61208-757-3

Automating Blog Crawling Using Pattern Recognition

Anal Kanti Roy[1], Nitin Agarwal[2]
Department of Information Science
University of Arkansas at Little Rock
Little Rock, Arkansas, USA
e-mail: [1]axroy@ualr.edu, [2]nxagarwal@ualr.edu

***Abstract*—Social media plays an important role in the propagation and dissemination of ideas and thoughts leading to the formation of diverse online communities. Compared to a myriad of other social media sites and applications, blogs provide a convenient platform for users to post detailed information, engage in active discussions and share the content on other social media sites, such as Facebook and Twitter. Thus, the blogosphere has been an enormous and ever-growing part of the open-source intelligence. In order to track and monitor online social behavior particularly from blogs, the first challenging part is to mine the vast pool of unstructured data. Several approaches have been developed to extract blog data using focused crawling, which requires a lot of time, effort and manual intervention. To scale up this process and cope with the continuously changing blog structure, we propose a sophisticated, advanced, generic, and scalable automated blog-crawler, with ability to identify different patterns in the Hypertext Markup Language (HTML) structure of the blog pages and extract data, such as title, author, date, content, tags, etc. from different blog posts. Using the crawler, we have crawled 530 blog sites with 894,856 blog posts so far.**

Keywords- blog crawling; generic crawler; blogs; blog posts; metadata; title; author; date; content; patterns; html.

I. INTRODUCTION

Recent years have witnessed an explosive growth of social media driven communications. Due to the pervasive nature of social media platforms, people have become more engaged in their ways of expressing thoughts. Usage of social media is no longer limited to just networking, advertising or self-expression. Sometimes, the negative side of the social platforms are encouraging people to utilize these online platforms for malicious purposes, such as spreading rumors, propaganda, polarization, extremism and radicalization.

The blogosphere (the clustered network of blogs and comments in existence on the internet and their links to other social media platforms) [1] is considered a highly dynamic subset of the social media. Starting from 1994, from early blogging platforms like LiveJournal, TypePad, and Blogger, the blogosphere has grown considerably with the addition of services like Tumblr, WordPress, Medium, Squarespace, etc. more than 500 million are recognized as blogs. Their authors account for over 2 million blog posts daily. Tumblr, possibly the biggest blogging platform, reports that it hosts over 440 million blogs. The most popular Content Management System (CMS), WordPress, adds about 60 million more [2]. Apart from the massiveness, this domain also distinguishes itself from the rest of social media platforms due to several features, such as more space for building discourse, more involvement in the conversation, and ability to spread into other platforms through shares. Analyzing blogs would therefore provide insights into our cyber behaviors, whether it is to monitor cyber campaigns, identify powerful actors and groups, study propaganda dissemination, and trace cyber threats [3][4]. However, collecting blog data is imperative for conducting any sort of computational analysis. This data collection process is quite challenging due to several reasons. First, the majority of information that can be crawled from blogs is unstructured and noisy, which makes it difficult to predict and model the crawling process. Second, the problem of automated crawling is exacerbated by the enormous growth rate of blog data.

To address the challenges mentioned above, we developed a generic crawler that is able to automatically parse the blog information for any blog site, categorize different data types based on their respective patterns, clean the data through data munging module and finally storing them in organized format in a Database Management System (DBMS). Although there is room for improvement for the automated crawler, it surely enhances the effectiveness and scalability of blog data collection. The rest of this paper is organized as follows. Section II describes similar blog crawling approaches taken by others so far. Section III explains separately how each type of blog data, such as title, author, date and content are extracted using our rule-based approach. The conclusion and acknowledgement close the article.

II. LITERATURE REVIEW

This literature review summarizes few blog-specific crawling methods available on the web and their drawbacks.

A. Web Content Extractor (WCE)

WCE is a crawling tool designed to extract web's data in general. For the extraction of data from the blogs, the patterns of the HTML pages and the patterns of the different data entities (title, author, date, etc.) has to be manually fed into the WCE before crawling any blog(s). Although WCE is quite accurate and can be operated without any prior knowledge of programming, but it has to be distinctly set up for every individual blog-site, which certainly is not a scalable solution when targeting multiple blogs. WCE has the feature of exporting the data into different formats .xml, .csv, .txt etc.) [5].

 ISBN: 978-1-61208-757-3

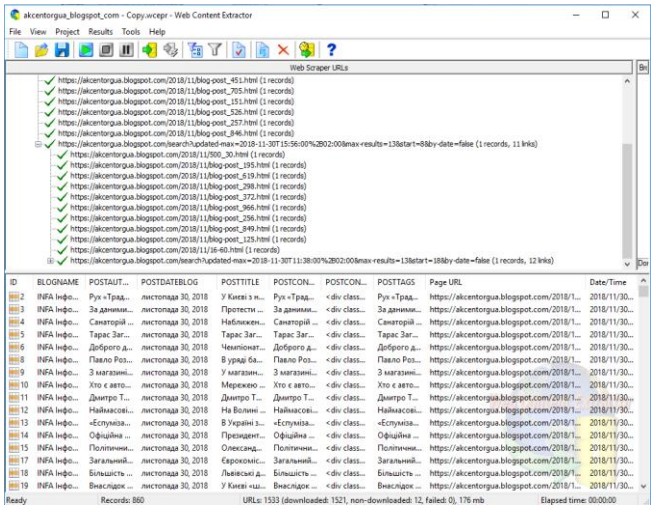

Figure 1. Extraction of blog elements by Web Content Extractor [5]

However, it does not have a synchronization ability with the database management systems. Therefore, a user has to convert the exported files manually into a compatible format that suits the database. It is also a multi-threaded web crawler that supports the data extraction from up to 20 threads (webs) simultaneously (Figure. 1 Extraction of blog elements by Web Content Extractor). However, the speed of the crawling process is normal as general crawler and sometimes slow while executing huge scripts. Moreover, this tool is completely focused and bound to the limited tasks with less room for customizations [5]. Therefore, WCE is quite not suitable to collect enormous amounts of data from a huge number of blogs.

B. Mapping the blogosphere towards a universal and scalable Blog-Crawler

The crawler in this method identifies the posts by crawling only feeds, RSS and atom from a host identified as a blog [6]. The crawler identifies the blog by downloading the crawled webpage and then parse it to check for the common standard patterns of the diverse blog systems. For instance, a pattern could be a match, if the generator tag of a web page contains < meta content = ‘‘blogger’’ name = ‘‘generator’’ />. Then crawler downloads the first alternate link rel = ‘‘alternate’’ with the type of an eXtensible Markup Language (XML) page recognized by type = ‘‘application / RSS + xml” or type=”application/atom+xml” and stores the referenced feed as the main feed of a blog, which should contain all posts. In some cases, the crawler may end up crawling the feeds, which contains only a minimized version of all posts, which are displayed as summaries on the home page of a blog(s). These instances occur because few web pages are limited to only displaying the latest posts. Consequently, the crawler has a drawback of crawling only the subsets of the displayed posts. The uncovered posts create a historical gap, as the crawler is unable to dive deep into the archives [6].

C. A New Algorithm of Blog-oriented Crawler

This work aims at downloading blog pages in some portal websites and the crawler views the blog as the special "topic" [7]. The concept of this method is identical to the topical crawling. Generally, topic describes the contents of a web page and is defined by the users before being fed as an input into the crawler. Similarly, for a blog as a topic, it refers to the types of blog pages and describes the structural patterns of those pages. It also can differentiate whether the fetched pages are relevant to predefined topic or not, and then the pages are downloaded if they are relevant, otherwise dismissed. The strategy of the topical crawler is to check for the contents of pages or structure of linkages. It prunes the URLs when crawling to some extends and orders the handled URLs. This crawler is neither a depth-first search nor a breadth-first search, but the best-first search. That means the crawler downloads the most relevant page, which the URL points to, in a current situation. This method also have a mechanism to extract blog linkages. However, it has some drawbacks in terms of precision and efficiency because of its generality and no specialty in nature [7].

D. BlogPulse

BlogPulse [8][9][10] devised by Intelliseek, was designed to find trends and patterns across selected 22,000 weblogs. It provided a list of key phrases, key persons and key paragraphs every day. It also had several analytical features that included gathering the insights by tracing the daily activities from the blogs, tracking the information by studying how the topic of information was disseminated through the blog posts. It profiled the top blogs, which provided detailed information about how the blog influences others and their activities. It extracted the list of 22,000+ weblogs from the BlogStreet directory. These blogs were crawled every day and were checked for possible duplication. The major limitation with BlogPulse is the restriction on the number of blogs available for crawling. Another challenge with the approach was that it was not user-friendly. Users were not allowed to monitor some set of blog sites and gather the trend information. Hence, the tool was a one-way reporting tool. It also does not fetch the comments from the blog posts, while crawling and mining the data. This tool was deprecated in 2012 [9].

Therefore, blog(s) oriented data retrieval systems are still in their premature stages. Blog-specific search engines only index the feeds, which usually are readable XML versions of the blogs. Most of them only provides a summary of the whole context obtained from the blog entries [11]. Some of them either focus on a tiny subset of blog(s) data with similar patterns [10], which in-turn consumes too much time and manual effort to crawl. The generic crawler we proposed allows us to crawl all the blog pages active in any blog irrespective of its different web structures saving significant amount of time and manual intervention which makes it different from other approaches.

 ISBN: 978-1-61208-757-3

III. METHODOLOGY

Since, the websites do not follow a unified standard that would allow the computer to semantically understand what every bit of HTML is trying to say, there is no common approach for grabbing patterns for every metadata. The ideal solution would be to build a robust and scalable tool that will be able to extract data irrespective of HTML content and structure. That is why; training the crawler for different HTML patterns of blogs, along with machine learning techniques should be an appropriate approach to apply. It will let the computer to understand what are on a HTML page exactly in the same way a human would.

First, we need an algorithm to recognize any blog pages. Among several papers [4][7][12] published, "Blog identification and splog detection" [12] suggests handful algorithms to identify blog pages by certain features like blog mark in URL, RSS tag, ordered dates in log etc.

A. Link extraction

After the recognition of blog pages, the next step is the extraction of all the relevant links (URLs) and identifying the patterns for required data. Here is the step-by-step approach for link extraction.

- Start crawling the URLs in every page of a blog site using Scrapy [13] that crawls in a DFO order.
- Filter out external URLs using Scrapy's rule-based approach [13], which only look for the domain name of the blog we are crawling (1).
  rules = (Rule (LinkExtractor (allow = [keyword], callback = 'parse_item', follow = True),) (1)
- Out of the internal URLs, it tries to eliminate the ones with a set of stop-words. Here goes the list:
  stopwords = ['facebook.com', 'google.com', 'twitter.com', 'youtube.com', 'pinterest.com', 'instagram.com', '/page/', '/search', '/author/', '/tag/', '/category/', '/about', '/comments/', '/contact/', '?', '=']

B. Extraction of Blog Content

Our method is to follow rule-based approach to recognize a specific information in a blog post. For example, 'title' of a blog post has a set of features or parameters that may classify it from any other piece of information. These features can be termed as classifiers. In order to filter out the post title, every chunk of data in the blog page has to go through the validation of these classifiers. Each classifier will provide them a score on a scale of 0 to 1 based on how they satisfy the conditions. The sum of scores from all the classifiers is the total score of each chunk of data. The chunk with highest score is considered as the post-title only if it surpasses a certain threshold. Since the importance or occurrence of all the classifiers is not the same, we multiply individual classifier scores by different weights before adding them. The weights can be determined by Naive Bayes Classification algorithm. We generated generic module for different data types, such as title, author, date and content. Here goes the explanation of step-by-step approach for each module.

C. Extraction of post title

Collect every chunk of text from the blog page by filtering out the empty nodes and store them in a list. Then perform required string manipulation in the chunks like stripping, joining and getting rid of unnecessary scripts. In this case, the classifier is set with the following patterns:

- The post titles are mostly likely to be surrounded by h1 or h2 tags.
- In blog site analysis, it can be observed that a post titles are not longer than 200 characters. Hence, this classifier filters out the descriptive chunks from the pool.
- The titles commonly appear at the beginning of the page. Therefore, the nodes with less depth in the HTML tree gets high priority.
- The post titles have a great chance to completely or partially match with the text between tags. The match percentage between these two texts could possibly lead to a decision. To match the similarity between the strings, we choose Cosine Similarity measurements here.
- In most of the cases, the URLs of the blog posts contain words quite similar to the post title delimited by '-' or '/'. Therefore, if we match the words of each chunk with the split words from the URL, the result significantly can contribute to a decision making.
- The post titles are often surrounded by tags with a class or id name of the "title", or a name where title is a substring. These scenarios can be used to identify the title quite easily.
- Most of the cases, the URLs of the blog posts contain words quite similar to the post title delimited by '-' or '/'. Therefore, if we match the words of each chunk with the split words from the URL, the result significantly can contribute to a decision making.
- The post titles do not usually end with a full stop. This filter out the other chunks from consideration.
- These individual scores are multiplied by their respective weights and then added up to obtain the total score.
- Equation (2) below shows that the chunk with highest score is extracted as post title only if it exceeds a certain threshold (3).

  Target = max (sum (individual scores * weight)) (2)

  Post title = TRUE if (Target > Threshold) (3)

- Now, the question is how the weights and the threshold are calculated. In this regard, we incorporate a data analysis over our collected feed

from a sample set of blog pages. Then, we use Naive Bayes classification algorithm.

TABLE I. CLASSIFIER TABLE

	Classifiers	True/False
Page 1	Between h1 tags	True
	Matches title tag text	True
	Class or id named title	True
Page 2	Between h1 tags	True
	Matches title tag text	True
	Class or id named title	False
Page 3	Between h1 tags	True
	Matches title tag text	False
	Class or id named title	False

TABLE II. FREQUENCY TABLE

Classifiers	True	False	Total
Between h1 tags	3	0	3
Matches title tag text	2	1	3
Class or id named title	1	3	3

TABLE III. LIKELIHOOD TABLE

Classifiers	True	Likelihood	Normalized-*3 (Weights)
Between h1 tags	3	1.00	3.00
Matches title tag text	2	0.67	2.00
Class or id named title	1	0.33	1.00

a. Rule-based classifiers for blog pages [14]

Let us assume that data are collected according to the performance of three classifiers from three blog pages where post title is taken as an example. Tables I, II and III show how weight is calculated. Likelihood refers to the chances of occurrence of a pattern. The pattern with more weight is given more significance while calculating the total score. From the above analysis, we calculate the threshold from the minimum total score that qualified as a post title, if the classifiers guessed correctly [14].

D. Extraction of post author

For any fetched URL, our method is to cross check with five predefined patterns, which identifies the author of a blog post. On passing the rules, each block of text gains a score just like post title extraction and then each score can be multiplied by their respective weights. Weights are determined by precision and recall of the training data. Finally, these scores are summed up for the highest scoring text block, which is considered as post author. However, the text block containing author name may contain other unnecessary texts like 'by', 'written by', 'and courtesy' etc. To filter these out, text-splitting methods, such as tokenization, n-grams and Inside–Outside–Beginning (IOB) tags from Natural Language Toolkit (NLTK) are used to pick human names.

Firstly, gather every chunk of text from the blog page filtering out the empty nodes and store them in a list. Then create a dataframe later on for storing records containing each block of text, its parent node and its score. For this purpose, initialize the following three lists and append values for texts and nodes for now. The scores are inserted after we have the calculations. Now, pass each block of text through the rules set by each pattern.

- The microformat rel="author" attribute in link tags (a) is commonly used to specify author of a post.

  Author

- Author related keywords are used in attributes like class and id etc. in the node containing author name. It can also be present in the href attribute if wrapped between anchor tags (a) like:

  href = http://www.example.com/author/name

- To capture these patterns, we do the following.

a) Firstly, we create a list of all the keywords that may possibly refer to author, such as 'author', 'byline', 'source', 'writer', 'written', 'by', 'courtesy', 'contributor', 'originator', 'creator', 'builder', 'editor' etc.

b) We then create a string joining all attribute values of the parent node for each text node.

c) Finally, we look for the existence of each of the author keywords in any of the attributes we gathered in the string above.

- Sometimes author names are specified in between meta author tags (<author>…</author>). Therefore, we check whether the parent node for each text is 'author' or not.
- The parent html tag for the author text is most likely to be anchor tags (<a>…). Therefore, this rule prioritizes the tag enclosed text blocks.
- The author information is supposed to be within a certain limit of text blocks. So, this rule emphasizes text blocks with a character limit of 50 or less in order to filter out larger chunks of text.
- In the end, all the scores multiplied by their respective weights are added up to yield the total score for each text block. The weights are calculated upon the precision and recall over the training data. We can use Bayes Classification algorithm or Random Forest model for this purpose. The more training data we have, the more it will lead to accuracy. For now, the weights are estimated based on a small set of data using Bayes algorithm. Finally, append score for each text block to the score's list we declared earlier.
- Create a pandas dataframe with lists (nodes, texts and scores) and select highest scoring text block as the post author data.

 ISBN: 978-1-61208-757-3

- The extraction is not finished yet. The text block containing author name may contain some unnecessary text along with the author name like 'by', 'written by', 'and courtesy' etc. To remove the surplus and extract author name only, we use tokens, n-grams, IOB tags, parts of speech etc. from NLTK, which can recognize human names most likely of a person or organization.

E. Extraction of post date

Alongside the traditional rule based approach, two additional approaches are used to safeguard the extraction of the 'date' of an article. Earlier approach used to traverse through the html body and recognize the highest ranked text chunk based on certain rules. If this approach does not work, it will look for a date in the 'content' attribute value of potential meta tags in html head section. If failed to find the date here also, it will finally try to pull out the date from the post URL, which is a commonly used trend to describe the resource path of a blog post or article. Even after successful extraction of postdate, it may contain unnecessary texts like 'On', 'Published on' etc., or the date may appear in a variety of formats sometimes in human readable forms like 'Yesterday', '2 mins ago', 'Tuesday' etc. So, the extracted data are finally processed through a date parser to get rid of unwanted texts and store the value in any user-required format (currently in 'dd-Month-yyyy'). Methods are described elaborately below along with code segments. The process of how normalized text chunks are fetched and ranking methodology that chooses the based suited chunk will remain the same. This section focuses more on the postdate patterns. Going through the source code will give a better understanding of the sequence.

This approach follows the traditional method of ranking each block of text, based on five patterns.

- The postdate commonly appears to be within the time tags (<time></time>). If crawler wants to fetch full date time format, it can find it in the 'datetime' attribute of the time tag.
- Date related keywords may be present in attributes like class, id, title, and content etc. in the node containing postdate. To capture this pattern, first create a list of all the keywords that may possibly refer to postdate. Then simply create a string joining all attribute values of the parent node for each text node. Finally, look for the existence of each of the date keywords in any of the attributes that are gathered in the string above.
- Now a days, it is common for resource path of a blog post to show date format in the post URL. As a first step, extract the date portion of the URL and match it with every text chunk. The higher the match ratio, the more possibility of that text to contain 'date of the blogpost'.
- There is a chance that a single blog post may have multiple dates. For example, each comment may contain datetime, which crawler is not looking for. So we emphasize on the original date value by its depth in the html tree as postdate, it usually appears at the beginning, most of the time after the post title.
- The postdates are no longer in character length. Therefore, we can filter out larger chunks of texts by limiting the character length of the text to 50 or less than equal to 50.

From the above process, crawler only looks for the post date in the html body section. If this approach fails to extract, it looks for the date in attributes of meta tags defined in the HTML head section, which most often stores the publishing date of the article/blog post. The following steps gives an idea about what meta tags, a crawler should look for and where the postdate may be present.

Out of various meta tags, only look for selective ones. Postdate usually appears in the meta tags, which contain these four attributes:

{Name, property, itemprop, http-equiv}

- Then look for potential keywords for the date in the values of these attributes. For each of the above attributes, a separate list of keywords are defined.
- If a match occurs, we take out the date information from 'content' attribute value of the same meta tag.

On failure of capturing date in spite of applying the above approach, we look for the date information in the URL of the article, which is a widely used fashion to define the resource path of an article. For example,

http://www.example.com/2017/05/29/blog-post-title

To do this, isolate the resource path from the URL through URL parsing and use regular expressions to partition the date block from it.

Even after successful extraction of postdate, the chunk may contain unnecessary texts like 'On', and 'Published on' etc., or the date may appear in a variety of formats sometimes in human readable forms like 'Yesterday', '2 mins ago', 'Tuesday' etc. The dateparser module does an excellent job to process the extracted date, get rid of unwanted texts and store the value in common format (currently in 'dd Month, yyyy').

F. Extraction of post content

Unlike post-title, post-author or post-date, it is quite challenging to extract post-content to an accurate satisfactory level by rule-based pattern identification approach. Rule-based approach results in inclusion of boilerplates and larger-sized comments by users. However, there are few resources in the web like Dragnet, Goose, Readability, Eatiht, Boilerpipe etc., which does the same kind of job of extracting main content. In this case, we can

 ISBN: 978-1-61208-757-3

use a tool "JusText" [15]. JusText is a useful tool to get rid of boilerplate content for example: navigation links, headers, footers and scripts from HTML pages. In a few cases, JusText cannot distinguish post-title, post-author and other unwanted texts and thus considers it as a part of the blog post/article-body. To avoid these, we can apply a few sanity checks and heuristics on the text returned by JusText. The procedure is explained step-by-step.

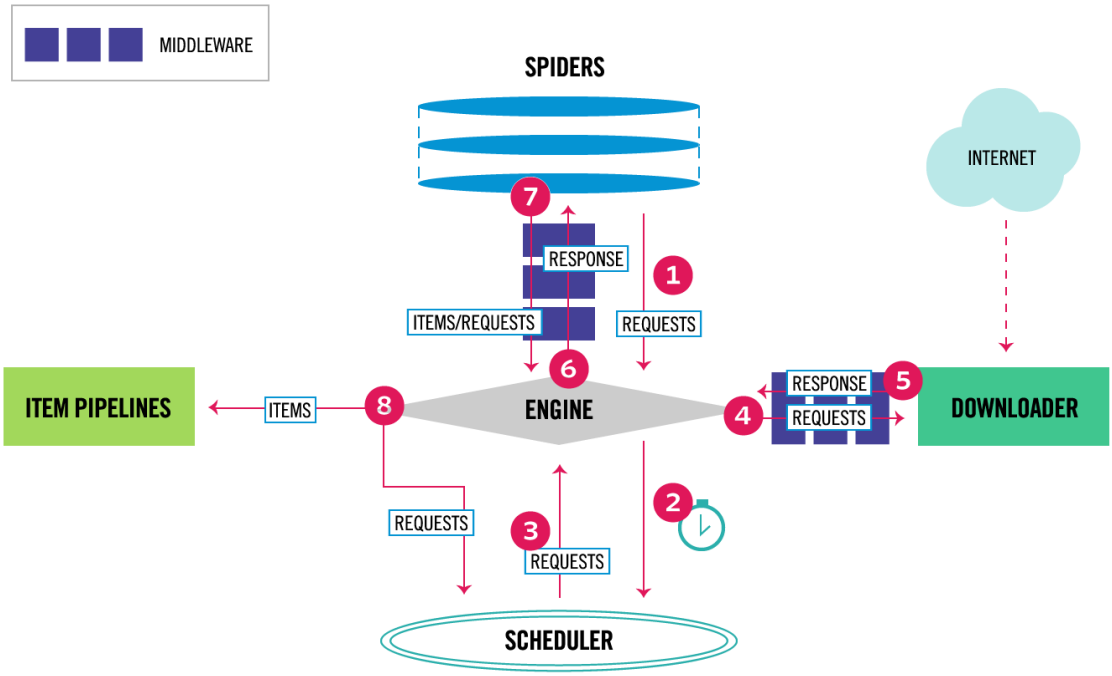

Figure 2: The architecture of crawling process by Scrapy [13]

- Justext can easily be installed via pip for either Python 2.6+ or 3.3+ [15].
- Then, requests can be sent to the blog-post URL to fetch pure texts from the response.
- Then comes the part of boilerplate detection. False cases in boilerplate detection are the texts we are targeting on, which further sanity-checks and heuristics should be performed.
- As mentioned earlier, these texts sometimes include other data like post-title, post-author or post-date along with the main blog-post content. So text-chunks can be checked to exclude these unnecessary data. By the time post-title, post-author and postdate are extracted, we can easily filter post-content out.
- Main blog-post contents are most likely to contain complete meaningful sentences. One tricky option to identify a complete sentence could be the text chunk ending with these punctuation marks: period (.), question mark (?) and exclamation point (!). This way unsolicited text chunks could be eliminated. We can also make sure the text chunks are trimmed on both right and left sides to ignore the dilemma of empty spaces.
- In very few cases, scripts are found embedded in the article body. These unnecessary scripts can be trimmed from the content.

The post content can also be extracted if the tag attributes of the html body for content contains the following property names in html body of the post.

['entry-content', 'article-body', 'article', 'articlebody', 'article-content', 'article-entry', 'post-body', 'post-entry', 'post-content', 'main-content', 'content', 'single-post', 'content-single', 'single-content', 'post', 'inner-content', 'page-content', 'postbody', 'material', 'material-body', 'article-main-content', 'material-body', 'materialbody', 'inner-post-entry', 'material-content', 'b-material content', 'article-inner-content']

G. Tools used

The tools used for the implementation are as follows:

- Scrapy: An open source and collaborative framework built on Python for extracting the data from websites in a fast, simple way and yet extensible by design. It allows us to plug new functionality easily without having to touch the core [13]. The diagram above (Figure. 2 The architecture of crawling process by Scrapy) interprets the architecture of the crawling process.
- Python: Python programming language (V. 3.7.1)

 ISBN: 978-1-61208-757-3

- MySQL Workbench: MySQL Workbench is a unified visual tool for database architects, developers, and DBAs.
- Elasticsearch: Elasticsearch is a search engine based on the Lucene library. It provides a distributed, multitenant-capable full-text search engine. It is developed in Java.

IV. EXPERIMENT

For experiment, we matched the data we extracted through our generic crawler with the respective sample data crawled by Web Content Extraction (expected to be correct). We collected data from 10 following blogs with 26,579 posts.

- http://europeans101.blogspot.com/
- http://ukrainianlaw.blogspot.com/
- http://www.asianpolicy.press/
- http://www.rabble.ca/
- https://futuristrendcast.wordpress.com/
- https://informnapalm.org/
- https://uprootedpalestinians.wordpress.com/
- https://www.asia-pacificresearch.com/
- https://www.counterpunch.org/
- https://www.no-to-nato.org/

Here are the overall experiment results based on averages.
Precision: 0.9412558436393738
Recall: 0.9580803573131561
F-measure: 0.9489225566387176

V. CONCLUSION

To sum up, in this paper we introduced the parsing of blog post pages of a blog for post attributes, using patterns that are automatically extracted from the blog's html patterns and implement a generic automated crawler for fast, robust and efficient blog data collection. We are still researching in what patterns, and to what extent blogs are interconnected. We also have great interest in analyzing the content of single weblogs. Due to this dynamic nature of blogs, we will face the long-term challenge of mining the blogosphere on a global scale. Even though the original implementation performed well along the milestones defined in the current crawler implementation, the accuracy of the crawler might not be 100%, but it can be smarter gradually adding more patterns to the data identification and by implementing machine-learning techniques.

ACKNOWLEDGMENT

This research is funded in part by the U.S. National Science Foundation (IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2605, N00014-17-1-2675, N00014-19-1-2336), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, and the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support. We also thank Mainuddin Shaik and Muhammad Nihal Hussain for valuable suggestions.

REFERENCES

[1] C. Fieseler, M. Fleck, and M. Meckel, "Corporate social responsibility in the blogosphere", Journal of business ethics, 2010 Feb 1, 91(4):599-614.

[2] Hosting Trubunal, "How many blogs are there" [Online], Available at: https://hostingtribunal.com/blog/how-many-blogs/, Last accessed on: Oct 6, 2019.

[3] M. N. Hussain, A. Obadimu, K. K. Bandeli, M. Nooman, S. Al-khateeb, and N. Agarwal, "A framework for blog data collection: challenges and opportunities", The IARIA international symposium on designing, validating and using datasets (DATASETS 2017), Jun. 2017.

[4] B. Mahar and C. K. Jha, "A Comparative Study on Web Crawling for searching Hidden Web", International Journal of Computer Science and Information Technologies 6.3 (2015), 1-5.K. Elissa, "Title of paper if known," unpublished.

[5] Newprosoft, "Web Content Extractor" [Online], Available at: http://www.newprosoft.com/web-content-extractor.htm, Last accessed on: Oct. 8, 2019.

[6] P. Berger, P. Hennig, J. Bross, and C. Meinel, 2011, "Mapping the Blogosphere--Towards a universal and scalable Blog-Crawler", 2011, IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, IEEE, Oct. 2011.

[7] L. Wei-jiang, R. Hua-suo, H. Kun, and L. Jia, "A new algorithm of blog-oriented crawler", 2009 International Forum on Computer Science-Technology and Applications, Vol. 1, IEEE, Dec. 2009.

[8] BlogPulse, Wikipedia [online], Available at: https://en.wikipedia.org/wiki/BlogPulse. Last Accessed on: Oct. 8, 2019.

[9] N. Glance, M. Hurst, and T. Tomokiyo, "Blogpulse: Automated trend discovery for weblogs", WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics, Vol. 2004, May 2004.

[10] L. Baker, "BlogPulse Search Engine" [Online], Available at: https://www.searchenginejournal.com/blogpulse-search-engine-launched-by-intelliseek/549/, May 2004, Last accessed on: 6 Oct, 2019.

[11] M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky (Eds.), "Advances in Information Retrieval: 28th European Conference on IR Research", ECIR 2006, London, UK, April 10-12, 2006, Proceedings. Vol. 3936. Springer, Mar. 2006.

[12] P. Kolari, T. Finin, and A. Joshi, "SVMs for the blogosphere: Blog identification and splog detection", AAAI spring symposium on computational approaches to analysing weblogs, Mar. 2006.

[13] Scrapy [online], Available at: https://scrapy.org/, Last Accessed on: Oct. 6, 2019.

[14] Naive Bayes Classifiers [online], Available at: https://www.geeksforgeeks.org/naive-bayes-classifiers/, Last Accessed on: Oct. 6, 2019.

[15] "jusText 2.2.0 - Heuristic based boilerplate removal tool" [online], Available at: https://pypi.org/project/jusText/, Last Accessed on: Oct. 6, 2019.

 ISBN: 978-1-61208-757-3

The Impact of Text Information Readability of Listed Companies' Annual Reports on Investors' Perception and Decision-making Behavior

Jinping Gao
Economics and Management School
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: doc.gao@126.com

Qin Xiao
Economics and Management School
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: 939865798@qq.com

Yan Wan
Economics and Management School
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: wanyan@bupt.edu.cn

Li Gao
School of Business and Management
Shanghai International Studies University
Shanghai, China
e-mail: li.gao@hotmail.com

***Abstract*—Readability is one of the most important characteristics of text information in the listed company's annual report. There exists a "framing effect" when decision makers use information, which means individual decisions are influenced not only by the content, but also by the way in which information is expressed. Readability greatly affects investors' trust perception and emotional experience, then, affects their decision-making. Based on the heuristic model of emotion and dual processing theory, this paper established a theoretical model to analyze the impact of readability on investors' trust perception, emotional experience and investment intention under different financial conditions. By using behavioral methods, we designed a 2 (readability: high, low) *2 (financial status: good, spread) mixed cross experiment to collect data. Our study reveals that regardless of profit status, investors are more willing to invest in companies with more readable annual reports and that the influencing process is mediated by investors' emotional experience and trust perception. These findings provided evidence for how the information disclosed by enterprises affects investors' perception and decision-making behaviors and also established theoretical basis for improving corporate information disclosure strategies.**

Keywords-annual report; readability; trust perception; emotional experience; investment intention.

I. INTRODUCTION

In the era of big data, information has increased explosively. Regulators and scholars increasingly emphasize the treatability of information. As early as 1998, the U.S. Securities and Exchange Commission (SEC) published the Plain English Handbook to improve the readability of the documents disclosed by enterprises. The China Securities Regulatory Commission (CSRC) has also emphasized that annual reports of listed companies can be illustrated with statistical charts and pictures so as to improve the readability. This study focused on the impact of readability on investors' trust perception, emotional experience and investment intention and we further analyzed the mediating effects of emotional experience and trust perception on investment intention.

As an important channel for listed companies to disclose information, an annual report contains financial information and non-financial information. Financial information is disclosed through standardized financial statements and will be audited by certified public accountants. Non-financial information is mainly disclosed through narrative text information, which provides useful incremental information for stake-holders [1]. Because narrative text information does not need to be audited, it may be easier for managers to manipulate to achieve impression management [2]. For example, managers may manipulate the readability of annual reports to obfuscate a company's performance [3].

Previous studies have shown that there are differences in the readability of management's information disclosure and this difference in readability is positively correlated with the market's short-term response [4]-[7]. Annual reports with low readability often lead to lower overall trading volumes and lower trading consensus [6], and investors are more likely to invest in companies with more readable disclosure documents [4]. On the contrary, annual reports with high readability are positively correlated with the trading volume of general investors, extraordinary returns and the possibility of corporate equity refinancing [5].

These results provide strong evidence for our study that the readability of disclosure documents can affect investment decisions and market results. However, previous studies only focused on the economic consequences of readability. They did not explore the causes of these results. Why did readability affect investors' behavior? The mechanism of the impact of readability on investors' decision-making remains to be further explored. Therefore, we introduced emotion and trust in psychology, trying to uncover the "black box" behind this

 ISBN: 978-1-61208-757-3

problem through the heuristic model of emotion and dual processing theory. This study not only provided new evidence for how information disclosed by enterprises affects investors' perception and decision-making behaviors, but also established a theoretical basis for improving corporate information disclosure strategies.

The rest of paper is organized as follows. Section 2 summarizes the literature review and develops our hypotheses. Section 3 describes the methods. Section 4 reveals the results and the analysis. Section 5 includes conclusions and suggestions.

II. LITERATURE REVIEW AND HYPOTHESES

Four hypotheses were proposed based on the literature on readability of annual reports, investment intention, emotion and trust.

A. Readability and Investment Intention

Readability is the cornerstone of text communication. It is the key factor which determines whether a text can be effectively understood [8]. According to Ease of Processing theory, people seem to prefer information which can be processed more easily [9]. Therefore, the more readable the information is, the smoother the process will be for investors, which also implies that such information disclosure is more reliable. Previous studies have proved that text readability can significantly influence the judgment of small investors [10][11]. Low readability tends to reduce investors' trust in information sources, leading to participants' poor evaluation [12]. Investors are more likely to invest in companies with more readable annual reports [5]. At the same time, readability often reflects executives' psychological characteristics and behavioral motivation to the capital market [13], and then affects the response of capital markets. Therefore, we believe that higher readability conveys more positive signals to investors and enhances their confidence and cooperation intention, thereby improving their investment intention. Thus, we propose the following hypothesis:

H1: Investors prefer to invest in companies with more readable annual reports.

B. Mediating Effects of Emotional Experience

In uncertain environment, individuals often use emotional heuristics to make choices. They consciously or unconsciously make decisions based on their subjective emotional reactions to task options [14]. Modern theories of cognitive psychology and neuroscience hold that there are two systems in human’s brain in the process of reasoning and decision-making, namely, the intuitive system and the rational system. The systems operate in parallel and depend on each other's guidance [15]. Both the heuristic model and the dual processing theory emphasize the role of emotional and intuitive irrational factors in decision-making. Analytical reasoning cannot be effective unless guided by emotion. Therefore, rational decision-making requires a proper combination of analytical system and intuitive system [15].

When investors make decisions, the company's business performance and future prospects provide important reference and will ultimately affect their investment intention and decision-making. This process is the result of careful analysis processed by the brain's rational systems. Rational analysis is important. But in uncertain and complex decision environment, people tend to rely on emotional experiences to make decisions [16]. A growing body of research has provided evidence that both positive and negative emotion can influence investor behaviors. Positive emotion will increase investors' risk tolerance [17]. In this case, investors will overestimate the return of investment decisions and underestimate the corresponding risks, thereby improving the level of investment [18]. However, negative emotion will increase the degree of investors' "loss avoidance" [17] and people will make careful analysis and critical evaluation when reasoning and making decision. Based on the above analysis, we predict that high readability will stimulate positive emotion and make investors enhance their investment intention and that low readability will generate negative emotion and make investors reduce their investment willingness. Therefore, we propose the following hypotheses:

H2: The impact of annual reports' different readability on investment intention is mediated by positive emotion triggered by investors. Annual reports with high readability will generate positive emotion and enhance the investment intention.

H3: The impact of annual reports' different readability on investment intention is mediated by negative emotion triggered by investors. Annual reports with low readability will generate negative emotion and reduce the investment intention.

C. Mediating Effects of Trust Perception

According to social psychology, trust is the process that the trustor generates confident and positive expected cognition based on words, behaviors and decisions of the trustee. In uncertain and uncontrollable circumstances, individuals tend to make judgments and decisions according to their trust perception. Trust Decision-making Model further explains the generation of trust that based on the trustor's perception of trustee's ability, integrity and kindness, the trustor will form a cognition of the trustee's credibility, and further form a trust evaluation for the trustee, thereby making the choice of whether to trust the trustee [19]. In the process of investment decision-making, an investor's first impression comes from the course of information collection and evaluation of listed companies. The investor will form a trust perception in listed companies and their management through reading or other intuitive feelings, and then make investment decisions. If the annual report's high readability increases the investors' trust perception in ability, integrity and kindness of the listed company and its management, and improves the investors' favorable impression for the company, we expect that investors will enhance their willingness to invest. On the contrary, the annual report's low readability may reduce investors' trust evaluation, thereby reducing their willingness to invest. Base on the analysis above, we propose the following hypothesis:

 ISBN: 978-1-61208-757-3

H4: The impact of annual reports' different readability on investment intention is mediated by the trust perception of investors.

III. METHODS

In order to verify the above four hypotheses, the required data were collected through experimental methods, including the design of participants, task materials, and procedures.

A. Participants

Participants were 40 graduate students of Beijing University of Posts and Telecommunications, including 10 males and 30 females. The age of participants ranged from 20 to 25. All participants had financial foundation and stock investment experience. Finally, a total of 38 valid data (9 males and 29 females) were collected.

B. Task Materials

Task materials were four simplified annual reports of different companies. Based on real annual reports disclosed by Shanghai Stock Exchange, we prepared two simplified annual reports according to the company's financial position (makes a profit or loss). On this basis, we changed their readability by changing the length of sentences and paragraphs, as well as altering information's present format (highlight the title and the main data, for example). By doing so, we controlled the consistency of the information content contained in two annual reports. Finally, we got 2 (readability: high, low) *2 (financial status: good, spread) totally 4 different company' simplified annual reports. Among them, company A and B were the profit group, in which A's annual report had high readability while B had low readability; C and D were the loss group, in which C's annual report had high readability while D had low readability.

In addition, some seven-point Likert-type scales were designed for our experiment. These scales included a reading feedback scale, an emotional scale, a trust scale, an investment intention scale and a basic information questionnaire. In the reading feedback scale, 4 items were designed to test how carefully the participants read the task materials. For the emotional scale, we used the Positive Affect and Negative Affect Schedule designed by Watson et al. [20] to evaluate investors' emotional experience. In our trust scale, we developed 12 items to evaluate the participants' trust perception on ability, integrity and kindness of the listed company and its management. In the investment intention scale, we designed 3 items to evaluate the investment intention of investors.

C. Task Procedures

We simulated the real investment environment and divided the tasks into profit group (A&B) and loss group(C&D). Each participant was asked to participate in both groups randomly. The main steps are as follows: participants were asked to read the annual reports of the two companies corresponding to the first task group. After reading, they needed to fill in the reading feedback and investment intention scale (the order of the two companies were random). After the completion of the first task, participants could rest freely and then followed the same process to complete the second task. After finishing both tasks, participants filled in the emotional scale, the trust scale and the basic information questionnaire. The experiment ended.

IV. RESULTS AND ANALYSIS

Paired-Samples T Test and Bootstrap mediation test were performed on the experimental data to test the validity of the above four hypotheses.

A. Readability and Investment Intention

According to the results of Paired-Samples T Test, readability had significant influence on investors' investment intention. As Figure 1 shows, the average investment intention of the company with more readable annual report is significantly higher than that of the company with less readable annual report in both groups (profit group: t=5.598, p=0.000; loss group: t=3.391, p=0.002).

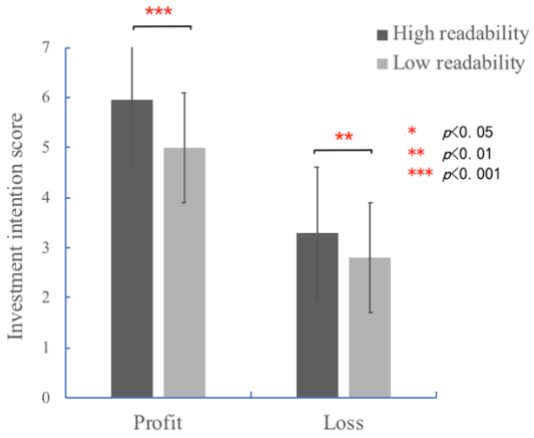

Figure 1. Investment intention under different surpluses

Therefore, H1 is supported. Regardless of profit status, investors prefer to invest in companies with more readable annual reports.

B. Mediating Effects of Emotional Experience

The results of Paired-Samples T Test showed that annual reports' readability causes significant differences in positive and negative emotion. Specifically, positive emotion triggered by high readability is significantly different from those triggered by low readability (profit group: t=7.712138, p=0.000; loss group: t=8.140316, p=0.000); negative emotion triggered by high readability is also significantly different from those triggered by low readability (profit group: t=4.091056, p=0.000; loss group: t=5.044795, p=0.000).

To further examine the mediating effects of emotional experience, we performed Bootstrap mediation test with reference to the mediation effect analysis procedure proposed by Zhao et al. [21] and multiple parallel mediation variable test methods proposed by Preacher and Hayes [22]. The sample size was selected as 5000 and the confidence interval was set as 95%. Table 1 shows the results.

 ISBN: 978-1-61208-757-3

TABLE I. MEDIATING EFFECTS OF EMOTIONAL EXPERIENCE

Status	Path	Effect	Boot SE	Boot LLCL	Boot ULCI
profit	direct effect	0.428	0.238	-0.047	0.903
	indirect effect	0.511	0.191	0.179	0.932
	positive emotion	0.462	0.245	0.014	0.956
	negative emotion	0.049	0.104	-0.143	0.277
loss	direct effect	-0.284	0.335	-0.953	0.385
	indirect effect	0.775	0.231	0.317	1.235
	positive emotion	0.474	0.216	-0.003	0.864
	negative emotion	0.301	0.153	0.052	0.660

Under the condition of profitability, because 0 was not contained in the interval (LLCI=0.179, ULCI=0.932), the two mediating variables, positive emotion and negative emotion, taken as a set, jointly mediated the relationship between readability and investment intention. The indirect effect through the two mediators was 0.511. An examination of the specific indirect effects indicated that only positive emotion was a mediator while negative emotion was not, since positive emotion's 95% CI did not contain 0 (LLCI=0.014, ULCI=0.956), and negative emotion's 95% CI contained 0 (LLCI=-0.143, ULCI=0.277).

The results indicated that, in profitable companies, the relationship between readability and investment intention was mediated by positive emotion, but not by negative emotion. Figure 2 shows the corresponding mediating model.

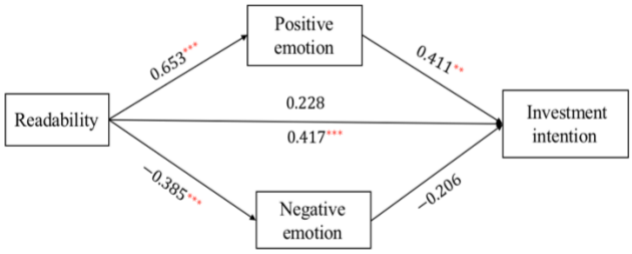

Figure 2. Emotional Experience mediating model of profitable company

Under the condition of financial loss, the interval (LLCI=0.317, ULCI=1.235) did not contain 0, indicating that the mediating effects of positive emotion and negative emotion were significant. The indirect effect through the two mediators was 0.775. In the two mediating paths, only negative emotion was a mediator, since its 95% CI did not contain zero (LLCI=0.052, ULCI=0.660). The mediating effect of positive emotion was not significant (LLCI=-0.003, ULCI=0.864).

The results indicated that, in loss-making companies, the relationship between readability and investment intention was mediated by negative emotion, but not by positive emotion. Figure 3 shows the corresponding mediating model.

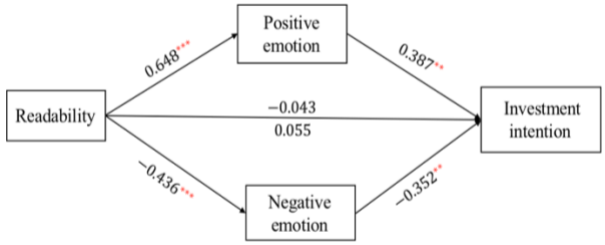

Figure 3. Emotional Experience mediating model of loss-making company

The interpretation of these results is that different emotions were at work under different surplus situations. In profitable companies, positive emotion mediated the effect of readability on investment intention. On the contrary, in loss-making companies, it was negative emotion but not positive emotion that mediated the effect of readability on investment intention.

C. *Mediating Effects of Trust Perception*

According to the results of Paired-Samples T Test, annual reports' readability caused significant differences in investors' trust perception (profit group: t=7.153509, p=0.000; loss group: t=5.821748, p=0.000). There were significant relations between trust perception and investment intention (profit group r=0.587, p=0.000; loss group r=0.415, p=0.001).

To further examine the mediating effect of trust perception, we conducted Bootstrap mediation effect test with reference to the mediation effect analysis procedure proposed by Zhao et al. [21] and Bootstrap method proposed by Preacher and Hayes [23] and Hayes [24]. The sample size was selected as 5000 and the confidence interval was set as 95%. Table 2 shows the results.

TABLE II. MEDIATING EFFECTS OF TRUST PERCEPTION

Status	Path	Effect	Boot SE	Boot LLCL	Boot ULCI
profit	direct effect	0.488	0.203	0.083	0.893
	indirect effect	0.450	0.137	0.218	0.756
loss	direct effect	0.048	0.283	-0.516	0.612
	indirect effect	0.443	0.158	0.171	0.790

Bootstrapping analysis indicated that 0 was not included in the indirect path regardless of profit status (profit group: LLCI=0.218, ULCI=0.756; loss group: LLCI=0.171, ULCI=0.790). These results revealed that trust perception mediated the relationship between readability and investment intention. H4 is supported. Figure 4 and Figure 5 show the corresponding mediating models.

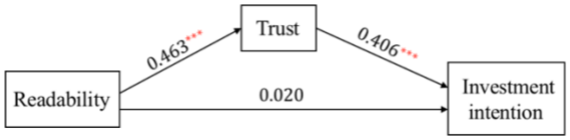

Figure 4. Trust mediating model of profitable company

 ISBN: 978-1-61208-757-3

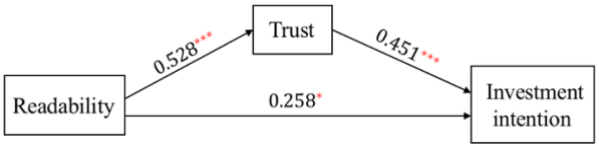

Figure 5. Trust mediating model of loss-making company

We used SPSS22 to analyze the above experimental data.

V. CONCLUSION

As one of the most important characteristics of text information in the listed company's annual report, readability considerably affected investors' decision-making. However, previous studies only focused on the economic consequences of readability. There were few empirical studies on the mechanism of readability' impact on investors. Based on the heuristic model of emotion and dual processing theory, this paper established a theoretical model to analyze the impact of readability on investors' trust perception, emotional experience and investment intention under different financial conditions. By using behavioral methods, we designed a 2 (readability: high, low) *2 (financial status: good, spread) mixed cross experiment to collect data. This study reveals that, regardless of profit status, investors are more willing to invest in companies with more readable annual reports and the influencing process is mediated by investors' emotional experience and trust perception.

This study provided new evidence for how information disclosed by enterprises affects investors' perception and decision-making behaviors. Firstly, we have provided evidence that investors prefer companies with more readable annual reports regardless of profit status. Secondly, we proved that the impact of readability on investment intention is mediated by emotional experience and trust perception. We interpret these results as that analytical system and intuitive system work together in the process of making judgments. When investors make decisions, the company's business performance and future prospects provide important reference and will ultimately affect their investment intention and decision-making. This is a process of careful analysis by the brain's rational systems. At the same time, the brain's intuitive system also plays a significant role that irrational factors, such as trust perception and emotional experience evoked by readability, affect investors' judgments and decisions as well. Our results are consistent with the notion that when investors make decisions, annual report information is processed by analytical system and intuitive system together and then affects investors' investment behavior.

This study also provided a theoretical basis for improving corporate information disclosure strategies. There exists a "framing effect" when decision makers use information, which means individual decisions are influenced not only by information content, but also by the way in which information is expressed. Readability affects investors' trust perception and emotional experience and then influences their decision-making behaviors. Therefore, in the process of information disclosure management, managers can make effective use of this influence by choosing more effective information disclosure methods (for example, disclose more readable documents) to send positive signals to investors and stakeholders, in order to improve the company's credibility and value identity and promote market and enterprise development more effective.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (Grant No.71372193). We thank Professor Y. Pan for helpful conversations and Dr. F. Wang for assistance with the task procedures.

REFERENCES

[1] M. Smith and R. J. Taffler, "The chairman's statement——A content analysis of discretionary narrative disclosures," Accounting Auditing & Accountability Journal, vol. 13, pp. 624-647, 2000.

[2] V. Beattie and M. J. Jones, "Measurement distortion of graphs in corporate reports: an experimental study," Accounting Auditing & Accountability Journal, vol. 15, pp. 546-564, 2004.

[3] R. Taffler and G. M. Smith, "The chairman's statement and corporate financial performance," Accounting & Finance, vol. 32, pp. 75-90, 2014.

[4] A. Lawrence, "Individual investors and financial disclosure," Journal of Accounting & Economics, vol. 56, pp. 130-147, 2013.

[5] T. Loughran and B. Mcdonald, "Measuring Readability in Financial Disclosures," Journal of Finance, vol. 69, pp. 1643–1671, 2014.

[6] B. P. Miller, "The Effects of Reporting Complexity on Small and Large Investor Trading," Accounting Review, vol. 85, pp. 2107-2143, 2010.

[7] R. Lehavy, F. Li, and K. Merkley, "The effect of annual report readability on analyst following and the properties of their earnings forecasts," Accounting Review, vol. 86, pp. 1087-1115, 2011.

[8] R. Rameezdeen and C. Rajapakse, "Contract interpretation: the impact of readability," Construction Management & Economics, vol. 25, pp. 729-737, 2007.

[9] R. Reber and N. Schwarz, "Effects of Perceptual Fluency on Judgments of Truth," Consciousness & Cognition, vol. 8, pp. 338-342, 1999.

[10] K. Rennekamp, "Processing Fluency and Investors' Reactions to Disclosure Readability," Journal of Accounting Research, vol. 50, pp. 1319–1354, 2012.

[11] H. T. Tan, E. Ying Wang, and B. Zhou, "How Does Readability Influence Investors' Judgments? Consistency of Benchmark Performance Matters," Accounting Review, vol. 90, pp. 371-393, 2015.

[12] A. L. Alter and D. M. Oppenheimer, "Easy on the mind, easy on the wallet: the roles of familiarity and processing fluency in valuation judgments," Psychonomic Bulletin & Review, vol. 15, pp. 985-990, 2008.

[13] X. Huang, S. H. Teoh, and Y. Zhang, "Tone Management," Accounting Review, vol. 89, pp. 1083-1113, 2014.

[14] P. Slovic, M. Finucane, E. Peters, D.G. Macgregor, and O. Azar, "Rational actors or rational fools: implications of the affect heuristic for behavioral economics," Journal of Socio-Economics, vol. 31, pp. 329-342, 2002.

[15] P. Slovic, M. L. Finucane, E. Peters, and D.G. Macgregor, "Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality," Risk Analysis, vol. 24, pp. 311-322, 2004.

 ISBN: 978-1-61208-757-3

[16] Kahneman, "Maps of bounded rationality: psychology for behavioral economics," The American Economic Review, vol. 93, pp. 1449-1475, 2003.

[17] D. Hirshleifer and T. Shumway, "Good Day Sunshine: Stock Returns and the Weather," Journal of Finance, vol. 58, pp. 1009-1032, 2003.

[18] U. Malmendier and G. Tate, "CEO Overconfidence and Corporate Investment," Journal of Finance, vol. 60, pp. 2661-2700, 2005.

[19] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," Academy of Management Review, vol. 20, pp. 709-734, 1995.

[20] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales," J Pers Soc Psychol, vol. 54, pp. 1063-1070, 1988.

[21] X. Zhao, J. G. Lynch, and Q. Chen, "Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis," Journal of Consumer Research, vol. 37, pp. 197-206, 2010.

[22] K. J. Preacher and A. F. Hayes, "Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models," Behavior Research Methods, vol. 40, pp. 879-891, 2008.

[23] K. J. Preacher and A. F. Hayes, "SPSS and SAS procedures for estimating indirect effects in simple mediation models," Behavior Research Methods, vol.36, pp. 717-731, 2004.

[24] A. Hayes, "Introduction to mediation, moderation, and conditional process analysis," Journal of Educational Measurement, vol. 51, pp. 335-337, 2013.

 ISBN: 978-1-61208-757-3

Predicting Opinions Across Multiple Issues in Large Scale Cyber Argumentation Using Collaborative Filtering and Viewpoint Correlation

Md Mahfuzer Rahman
University of Arkansas
Fayetteville, AR, USA
email: mmr014@uark.edu

Joseph W Sirrianni
University of Arkansas
Fayetteville, AR, USA
email: jwsirria@uark.edu

Xiaoqing (Frank) Liu
University of Arkansas
Fayetteville, AR, USA
email: frankliu@uark.edu

Douglas Adams
University of Arkansas
Fayetteville, AR, USA
email: djadams@uark.edu

Abstract—**One challenging problem in large-scale cyber argumentation is that discussions are often incomplete as some ideas only get addressed by a fraction of the users. Typically, users engage only with some ideas but not all of them, making it difficult to assess collective intelligence. To resolve this problem, we developed an innovative method of predicting a user's opinion on ideas that they have not discussed using the opinions from related ideas with intelligent argumentation and collaborative filtering. Our method considers the similarity of users and the correlation of different ideas across issues to make predictions. Compared to other existing opinion prediction methods, experimental results on an empirical dataset show that our method is 21.7% more accurate. Two major innovative contributions are made in this research: 1) We developed a novel approach to predict a participant's opinion on a non-participated idea using similar users' opinions from related ideas with an excellent accuracy in cyber argumentation; 2) This is the first research to enable multi issue opinion prediction with partial agreement on an idea. This is encouraging from several perspectives. This prediction model will help to assess collective intelligence from cyber argumentation more accurately by providing additional data both in individual and collective level. In addition, it may speed up a cyber argumentation analysis process by reducing the amount of participation required.**

Keywords-opinion prediction; incomplete ongoing discussion; collaborative filtering; cyber argumentation; collective intelligence.

I. INTRODUCTION

In large-scale cyber argumentation platforms, participants express their opinions, engage with one another and respond to feedback and criticism from others in discussing important issues online. Cyber argumentation platforms implement argumentation models to enforce an explicit discussion structure, such as Dung abstract frameworks [1], Issue-Based Information Systems (IBIS) [2], and Toulmin's model of argumentation [3]. These structures allow argumentation analysis tools to effectively analyze the discussions. Argumentation analysis tools can capture the collective intelligence of the participants and reveal hidden insights from the underlying discussions. In this research domain, these tools have demonstrated the ability to evaluate and reveal hidden phenomena, such as identifying group-think [4], polarization [5], assessing argument validity [1], etc.

However, such analysis requires that the issues have been thoroughly discussed and participant's opinions are clearly expressed and understood. Participants typically focus only on few ideas and leave others unacknowledged and under-discussed. This generates a limited dataset to work with resulting in an incomplete analysis of issues in the discussion. This also hampers the individual and collective intelligence retrieval process and opinion analysis from the underlying discussion. Particularly a limited dataset with missing values affects the clustering or user grouping algorithms and the resulting user groups introduce error and bias in different social phenomena analysis [6].

One solution to this problem would be to predict a participant's opinion with high accuracy on an idea that they have not explicitly expressed. With reasonably accurate prediction of missing information, we can analyze the individual and collective opinion of users effectively even if they did not participate in some of the discussion. Collective intelligence can also be assessed more accurately when discussions are incomplete. Predicted values can also fill the missing information for clustering algorithms and the derived group related analytical models.

In this paper, we present a method of predicting participant's opinions on different ideas that they have not explicitly engaged with. We use our argumentation platform, the Intelligent Cyber Argumentation System (ICAS), to collect user opinion on issues and predict the missing opinions. In our system, discussions take on a tree structure. Issues are the root of the conversation. Under an issue, there are a finite set of different positions that address the issue. We use a collaborative filtering model based on viewpoint correlation between positions and user opinion similarity to predict user's missing opinion on a position.

We compared our method Cosine Similarity with Correlation based Collaborative Filtering (CSCCF), with other opinion prediction methods based on popular predictive techniques on an empirical dataset collected with our argumentation platform, ICAS. Our dataset contains over ten thousand arguments on four issues and sixteen associated positions from more than three hundred participants. The experimental results show that our model has good accuracy and is 21.7% more accurate on average than other benchmarking methods.

In this paper, we make the following contribution:

- We propose a model (CSCCF) for predicting user opinion on positions using collaborative filtering based on viewpoint correlation between positions and user opinion similarity.
- We compare our model with other popular predictive techniques on an empirical dataset and show that our method is more accurate.

 ISBN: 978-1-61208-757-3

- We demonstrate how our method is capable of predicting several different positions at once without significantly compromising accuracy.

The rest of the paper is structured in the following way. In Section 2, we discuss previous research works which are related in different aspects/ways with the work presented in this paper. In section 3, we give a brief description about our argumentation platform ICAS and how we derive user's opinion in different issues. Section 4 describes the CSCCF opinion prediction model to predict missing opinion values. In Section 5, we talk about the empirical study to collect dataset, and different experiments to evaluate our CSCCF model. The remaining sections contains the Discussion, Conclusion and Reference for this work.

II. RELATED WORK

This section describes previous research works which are related in four different aspects with our work presented in this paper.

A. Opinion Analysis on Argumentation Platform

Many researchers have worked on analyzing user opinion in cyber argumentation system, such as opinion space [7] and Considerit [8] etc. Their main objective was analyzing how users engage with different opinionated people or ideas and how it affects their overall opinion. These platforms mostly focused on analyzing collective user opinion from user participation data only. None of these platforms have attempted to predict user opinion on non-participated issues.

B. Opinion Prediction on Social Media

Social media data is often used by many researchers to work on collective user stance/opinion prediction. Political discussions on twitter have been used to classify user political stance [9]. Social media data was also used to predict user reaction on certain events, such as the 2015 Paris Terror Attack [10] or classify people's stance on important issues [11]. These works mostly looked at predicting opinion on a single issue using the related textual content on that issue only, they are not using the user opinion in related issues to infer opinion in another issue like our method presented in this paper.

C. Multi-Issue Opinion Prediction

Little work has been done on an individual's opinion prediction across multiple issues. [12] used Probabilistic Matrix Factorization (PMF) to fill out a user-aspect opinion matrix (aspects are analogous with issues) as an intermediate step of a larger process to predict the polarity of interaction between users. However, since this was an intermediate step, the authors did not evaluate the success of the prediction step. [13] used traditional collaborative filtering methods to predict user's opinion on important political topics. In a follow-up paper [14], they used topic distribution from user arguments, user interaction and profile data to infer a user's stance on an issue. In their system, each issue only had two positions and users can only agree or disagree with a position. Whereas in our system each issue can have multiple positions and user can agree or disagree with a level of agreement from -1.0 to +1.0.

D. Different Variation of Collaborative Filtering

One of the major differences between different memory based collaborative filtering (CF) algorithms is how they calculate similarity between users/items to predict missing values from the most similar users/items. One popular approach measures the correlation between two users/items and use it as a similarity measurement between them [15], such as Pearson Correlation, Kendall's τ correlation. Cosine similarity of two user/item vectors is also used to measure similarity among them [15]. To our knowledge there is no similarity method that uses correlation values of items as weight in cosine similarity measurement like our method.

Some CF approaches measure the correlation values between different data domains. Collective Link Prediction, and Multi-domain Collaborative Filtering [16] are some of the models which exploit domain correlation via different learning based methods. Collective or Relational Matrix factorization [17] models use correlation between multiple relations for relational learning when an entity/user participates in multiple relations. Cross domain CF model uses this approach via coordinate system transfer method [16]. However, these models are computationally expensive and used to figure out correlations in between different data domains or multiple relations. Whereas, our model exploits the correlation within one data domain or in a single relation between user and item in a computationally inexpensive way.

III. ICAS SYSTEM

We use our intelligent cyber argumentation platform ICAS to derive viewpoint vectors for each participant, which are later used for opinion prediction. ICAS is a cyber-argumentation platform that is capable of automatically determining the opinion of participants towards different positions in the discussion. ICAS is the enhanced version of the online argumentation system developed in prior work [18].

A. ICAS Architecture

In the ICAS architecture, discussions take on a tree structure, with issues at the root of the tree, positions solving/addressing the root issue on the first level of the tree, and the arguments made for or against the positions or other arguments in the position as the remaining nodes in the tree. Participants contribute to the discussion by making arguments. Arguments are statements of agreement (for or against) and rationale relating to their parent node. Arguments can be made to support/attack positions or refute/agree with other arguments. When writing an argument, participants fill out two fields. First is the argument text, where they give their rationale for the argument. The second is the level of agreement. Here, users choose their level of agreement on a weighted scale from -1.0 to +1.0 at 0.2 length intervals. The sign of the agreement level indicates whether the user is agreeing (positive) or disagreeing (negative) with the parent node. The magnitude of the agreement level indicates the intensity of the agreement, where a lower magnitude is closer to indifference and a greater magnitude is closer to complete agreement/disagreement. For example, an agreement level of +0.8 would represent a very high level of support, while an

 ISBN: 978-1-61208-757-3

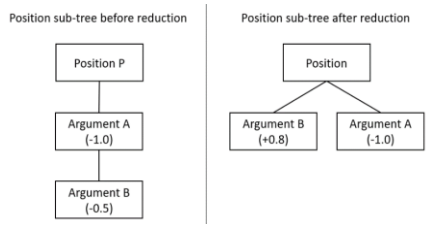

Figure 1. Example of an argument reduction. Argument B is reduced from the second level of the tree to the first level.

agreement level of -0.4 would represent a moderate level of disagreement.

B. Deriving Viewpoint Vectors using ICAS

A viewpoint vector is a vector where each element represents a user's opinion toward a position being discussed. We average the agreement values of the arguments a user posted under a position to determine a user's opinion toward the position. A user can post arguments supporting or attacking other user's arguments at different levels of argument tree. The associated agreement values state user's agreement with the parent argument, not with the root position directly. We used argument reduction method [19] to connect all these arguments to the root position. This method uses artificial intelligence, fuzzy logic, linguistic heuristic rules and other techniques to reduce an argument from any level of argument tree to the first level considering the support/ attack relationships with updated agreement value. The updated value represents the argument's agreement value directly towards the root position. Fig. 1 visualizes this reduction. For a more in-depth explanation of the fuzzy logic engine and argument reduction method, refer to [19, 20]. Argument reduction method is not 100% accurate, instead this is an estimation of user's opinion towards the root position. Several case studies have shown that this method achieves reasonable accuracy [19][20].

IV. Opinion Prediction Model

The section describes the CSCCF model for missing opinion value prediction. It is divided into three sub sections which describes required data for CSCCF model, steps and algorithms for CSCCF, and time complexity of CSCCF model.

A. Data Required for Prediction

To predict a missing opinion value at a position we need the following information: 1) the viewpoint vectors of all users in the training data, 2) opinion correlations of different positions with the target position t, and 3) the target user's viewpoint vector.

A viewpoint vector represents the opinions or agreement values of a particular user for all positions in the system. At training time, our model calculates the viewpoint vectors for every user. If there are n different positions on various issues in the system, we can represent a user's viewpoint vector in the following format:

$U_i = [R_1^i, R_2^i, R_3^i, R_4^i \ldots\ldots R_n^i]$; here U_i is the viewpoint vector for user i and R_p^i is the opinion value of the user i at position p. If user i did not participate in position p discussion, then R_p^i will be represented as invalid or missing value.

The correlation value between two positions indicates how much participant's opinions are associated in these two positions. A strong correlation value indicates if a user agreed in one position, whether the user agreed or disagreed in another position and vice versa. The correlation vector of a position can be formalized in the following format:

$C_p = [C_{p1}, C_{p2}, C_{p3}, \ldots, C_{pn}]$; here C_p is the correlation vector of position p and C_{pq} is the correlation between position p and position q, C_{pp} would represent the correlation value between the same position p, which is 1. Although this value will not be used in predicting position p, only the related positions with position p will be used. We calculated the correlation values between positions using the Pearson Correlation Coefficient from the training data and only considered the correlation values with high confidence (Two tailed p-values above 0.05 are discarded).

The target user's viewpoint vector can be represented in the following format:

$U_x = [R_1^x, R_2^x, R_3^x, R_4^x \ldots, R_{t-1}^x, ?, R_{t+1}^x \ldots\ldots R_n^x]$; Here, R_t^x, the value at position t is missing, we will predict this value.

B. Opinion Prediction using Cosine Similarity and Position Correlation

We want to predict the opinion value of user x at position t, the R_t^x value in U_x. This process has two steps. First, we need to identify the most similar users to user x with respect to position t from our training data. Second, we need to aggregate their opinion values at position t to use it as predicted value.

To identify the most similar users with respect to position t, we filter out the users who have a missing value at position t in their viewpoint vector. The remaining users are placed into user x's candidate set. Then the similarity between target user x and every user in the candidate set is calculated.

To calculate similarity between two users x and y, we first remove any elements from the vectors at which either vector has a missing value. Given, U_x and U_y are the viewpoint vectors of user x and users y. U_x has a missing value at position t, so we remove R_t^x and R_t^y from the vectors.

$U_x = [R_1^x, R_2^x \ldots R_{t-1}^x, R_{t+1}^x \ldots\ldots R_n^x]$

$U_y = [R_1^y, R_2^y \ldots R_{t-1}^y, R_{t+1}^y \ldots\ldots R_n^y]$

Next, the viewpoint vectors are updated using the values from target position's correlation vector, C_t. Each value in the viewpoint vector is multiplied by its corresponding position correlation value with target position t. The updated viewpoint vectors are represented as $U_x^\wedge$ and $U_y^\wedge$:

$U_x^\wedge = [C_{t,1}R_1^x, C_{t,2}R_2^x, .. , C_{t,t-1}R_{t-1}^x, C_{t,t+1}R_{t+1}^x, .., C_{t.n}R_n^x]$

$U_y^\wedge = [C_{t,1}R_1^y, C_{t,2}R_2^y, .. , C_{t,t-1}R_{t-1}^y, C_{t,t+1}R_{t+1}^y, .., C_{t.n}R_n^y]$; here, Opinion value at position i is multiplied by $C_{ti;}$ the correlation value between position i and t.

Then, we calculate the cosine similarity between the updated viewpoint vector $U_x^\wedge$ and $U_y^\wedge$ to determine how similar user x and y are with respect to position t using (1).

The similarity value lies in between [-1,1], where -1 represents complete difference, 0 represents no correlation, and 1 represents complete similarity.

Similarity (user x, user y) = Cosine Similarity ($U_x^{\wedge}$, $U_y^{\wedge}$)

$$= \frac{\sum_{i=1, i\neq t}^{n} C_{ti}^2 R_i^x R_i^y}{\sqrt{\sum_{i=1, i\neq t}^{n} C_{ti}^2 (R_i^X)^2} + \sqrt{\sum_{i=1, i\neq t}^{n} C_{ti}^2 (R_i^y)^2}} \quad (1)$$

Using the above method, we calculate the similarity between target user x and every user in x's candidate set. Then, we rank all the users based on their similarity value with target user x and select the top k neighbors, where k is a constant model parameter. We experimented with different values for k (3, 5, 10 etc.), we got the best result when k was set at 5 on the dataset we validated this model. The model then averages the opinion values of top k neighbors at position t weighted by the similarity value to predict the value of R_t^x as shown in (2).

$$\text{Predicted value of } R_t^x = \frac{\sum_{m=1}^{k} Similarity(x,m) * R_t^m}{\sum_{m=1}^{k} Similarity(x,m)} \quad (2)$$

Our method finds the most similar users to the target user with respect to the position we are predicting. Multiplying the opinion values with the associated test position correlation values weights the opinion values as per their importance to determine the test position. It also filters out the uncorrelated opinion values in similarity calculation.

C. Time Complexity of CSCCF Model

Let, number of users = n and number of positions = m, we will measure the time complexity to predict a missing opinion value for one test user. We calculate the correlation values between the positions from the training data only one time and use it to predict the missing opinions for all test users. To make one single prediction, first we calculate the cosine similarity between updated viewpoint vectors n times, one for each user. Then, we sort the similarity values from n users and make prediction from top k neighbors. The time complexity of these two steps are O(n*m) and the time complexity of sorting n numbers respectively. In our case, the time complexity of sorting n number was O(nlogn) as we used heap-based priority queue. So, the overall time complexity of our algorithm is O(n*m) + O(nlogn).

V. EXPERIMENTS

This section describes the empirical study, dataset collection process and experimental setup to evaluate our CSCCF model.

A. Empirical Data Description

We conducted an empirical study in spring of 2018 on a group of 344 undergraduate students in an entry level sociology class. The students were asked to discuss four issues, each with 4 different positions over the course of five weeks. The resulting discussion had over 10000 arguments, from 309 users. 90 out of 309 users had complete participation. On average 69 users (22.33%) had missing opinion values in the positions. We received Institutional Review Board (IRB) approval from the university to conduct this empirical study and use the anonymized data for research purposes. Table 1 describes the dataset with issues and positions.

B. Methods to Test Against

We tested our model (CSCCF) against following different popular predictive techniques to compare accuracy. The only difference between CSCCF and other CF based models is the way similarity between two users is measured.

1) Cosine Similarity based CF (CSCF) : This CF model used the Cosine similarity between the original viewpoint vectors U_x and U_y, to calculate similarity between user x and y using (3):

$$\text{Cosine Similarity } (U_x, U_y) = \frac{\sum_{i=1, i\neq t}^{n} R_i^x R_i^y}{\sqrt{\sum_{i=1, i\neq t}^{n} (R_i^X)^2} + \sqrt{\sum_{i=1, i\neq t}^{n} (R_i^y)^2}} \quad (3)$$

2) Neural Net : We implemented a neural net that uses hybrid latent variables as described in [21] to learn individual

TABLE I. DATA DESCRIPTION WITH ISSUES AND POSITIONS

Issue Name	Position No	Position Text
Guns on Campus: Should students with a concealed carry permit be allowed to carry guns on campus?	0	No, college campuses should not allow students to carry firearms under any circumstances.
	1	No, but those who receive special permission from the university should be allowed to concealed carry.
	2	Yes, but students should have to undergo additional training.
	3	Yes, and there should be no additional test. A concealed carry permit is enough to carry on campus.
Religion and Medicine: Should parents who believe in healing through prayer be allowed to forgo medical treatment for their child?	4	Yes, religious freedom should be respected.
	5	Yes, but only in cases where the child's life is not in immediate danger.
	6	No, but may deny preventative treatments like vaccines.
	7	No, the child's medical safety should come first.
Same Sex Couples and Adoption: Should same sex married couples be allowed to adopt children?	8	No, same sex couples should not be allowed to legally adopt children.
	9	No, but adoption should be allowed for blood relatives of the couple, such as nieces/nephews.
	10	Yes, but same sex couples should have special vetting to ensure that they can provide as much as a heterosexual couple.
	11	Yes, same sex couples should be treated the same as heterosexual couples and be allowed to adopt via the standard process.
Government and Healthcare: Should individuals be required by the government to have health insurance?	12	No, the government should not require health insurance.
	13	No, but the government should provide help paying for health insurance.
	14	Yes, the government should require health insurance and help pay for it, but uninsured individuals will have to pay a fine.
	15	Yes, the government should require health insurance and guarantee health coverage for everyone.

 ISBN: 978-1-61208-757-3

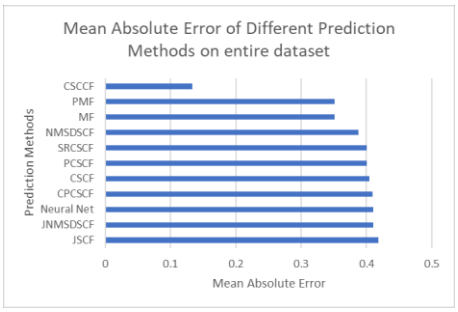

Figure 2. Mean Absolute Error of different Models on entire dataset

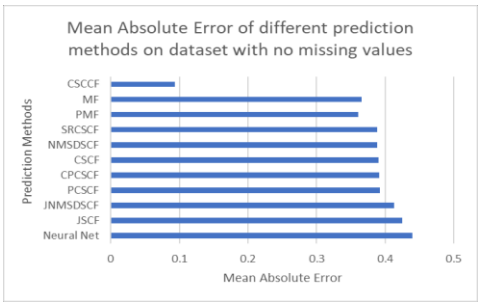

Figure 3. Mean Absolute Error of different Models with no missing values

information about both the users and positions. The neural net learns model weights and latent input variables during training. Various input latent vector sizes were tried, latent vectors with length 2 for both users and positions did best. The topology of the neural net is: linear layer(4, 6) => Tanh layer(6,6) => linear layer(6,1) => Tanh layer(1, 1). The first argument for the layer is the input size, the second is the output size. The neural net used stochastic gradient descent and optimized for sum squared error (SSE).

3) Matrix Factorization (MF) : We implemented Regularized Incremental Simultaneous MF as described in [22] which decomposes the user-position matrix (|U| * |D|) into two matrices (|U| * |K| and |D| * |K|) to discover K latent features. In oreder to avoid overfitting, this method applies regularization by penalizing the magnitude of vectors. It was optimized for SSE. We tried different sizes for latent factor K, the best result was found when K was 5.

4) Probabilistic Matrix Factorization (PMF) : We implemented PMF as described in [23]. The latent matrices are drawn from a gaussian distribution, determined by the means and variances of each row in the original user-position matrix and was optimized for SSE. Different latent factor sizes were tried and 5 did best for PMF.

5) Spearman Rank Correlation Similarity based Collaborative Filtering (SRCSCF): We ranked the original viewpoint vector (U_x and U_y) and measured the similarity between user x and y using (4):

$$Sim(user\ x, user\ y) = 1 - \frac{6\sum_{h=0}^{n} d_h^2}{n(n^2-1)} \quad (4)$$

Here, d_h is the difference in the ranks for item h by the user x and y, n is the number of co-rated items.

6) Pearson Correlation Similarity based Collaborative Filtering (PCSCF) : Pearson correlation coefficient value of U_x and U_y is used to measure similarity between users.

7) Constrained Pearson Correlation Similarity based Collaborative Filtering (CPCSCF) : This method uses midpoint instead of mean value from U_x and U_y in Pearson correlation to measure similarity between users.

8) Jaccard Similarity based Collaborative Filtering (JSCF) : We have rounded the opinion agreement values in U_x and U_y upto two decimal points and measured the Jaccard coefficient as similarity using (5):

$$Sim(user\ x, user\ y) = \frac{|U'_x \cap U'_y|}{|U'_x \cup U'_y|} \quad (5)$$

9) Normalized Mean Squared Difference Similarity based Collaborative Filtering (NMSDSCF): It uses the normalized mean squared difference (NMSD) of rating vectors as difference between users to calculate similarity.

10) Jaccard and Mean Squared Difference Similarity based Collaborative Filtering (JNMSDSCF) : This method multiplies similarity value from JSCF and NMSDCF to calculate similarity between users.

C. Results

We tested our model CSCCF along with other comparison models and measured the Mean Absolute Error (MAE) from the predicted and actual opinion value for the following experiments. We performed a cross validation with 5 fold and 2 repetitions and the data was separated as 80% training and 20% testing in each iteration.

1) Accuracy on entire dataset: We measured the MAE for each position separately and then averaged the results. Fig. 2 summarizes the result of this experiment. On average our model achieved a MAE value of 0.133. The second most accurate model, PMF achieved a MAE value of 0.350 and the other models were all in between 0.351 to 0.42. This shows that our model is a distinct improvement over other models. As most of the users did not participate in all positions, this dataset contains lots of missing information which is hampering other models. Our model handled this sparsity problem incorporating global correlation values from training data and used them as weight to prioritize the limited available opinion values in the similarity calculation.

2) Accuracy on dataset with no missing values: We also tested how the accuracy of different models would change if we only consider the data with no missing values. This dataset is much smaller as only 90 out of the 309 students

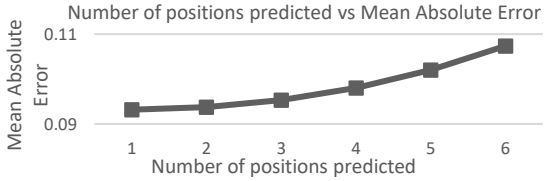

Figure 3. Number of positions prediction vs Mean Absolute Error

participated in all the position discussions. Fig. 3 summarizes the results of this experiment. MAE value of all the models tended to decrease for this dataset. On average our model's MAE decreased to 0.093 and for the second-best model, PMF it was 0.365. These results show that our model outperformed other models even on a complete dataset with no missing values. We think smaller dataset is the main reason of higher MAE values from neural net, and matrix factorization based models. And Prioritizing opinion values in similarity calculation is helping us to achieve lower MAE value than other CF based models.

3) Predicting multiple positions: We tested the accuracy of our model if we predict 2 to 6 positions simultaneously. At each number of predictions we considered all possible combination of position indices. As example, when we predict 2 positions at once, we tested with all 120 position combination of two position indices as testing positions and averaged the MAE values. Fig. 3 shows the result from this experiment on the dataset with no missing values. The MAE increases when more positions are being predicted at once. After 3 positions, the MAE increases at a faster rate but remains relatively low. The main cause of higher MAE is that we might be predicting correlated positions simultaneously. For example, if we are predicting correlated positions 1 and 3, position 3's value will not be used in postion 1's similarity calculation and vice versa.

VI. DISCUSSION

We think our model is working because people's opinions are correlated across different issues due to their similarity in terms of their values in the sense of Schwartz theory of basic human values [24]. Their political leanings, such as conservative, lean conservative, lean liberal, liberal etc. and their position on religion are few of the issues deriving from their values. Generally people choose a certain perspective on social issues based on their political leanings which our model captures using the correlation values between positions.

The improvement over CF models notably the CSCF shows the importance of using viewpoint correlations in opinion prediction. Each opinion value had the same priority in similarity calculation in these models whereas in our model opinion values were weighted according to its correlation with the test position. The improvement over the neural net, MF, and PMF methods is likely because of the limited data size. The latent features for the users and positions were probably underdeveloped and contained little meaningful information. If each user had more data points, then these models might have done better. There is no straightforward way to filter out uncorrelated positions in these models. Neural Net automatically figures out which features are irrelevant, but the lack of data is preventing it from doing it. In CF or MF, missing values are predicted on an initial user-item matrix, there is no common way to filter out different item set for different item predictions.

If there is a strong correlation between the ratings of different data items from the overall users, we think our CSCCF model will generate good prediction results. Also, it might help to deal with the cold start and sparsity problem especially when a user has provided very few opinions on related issues. In order to achieve high accuracy by our CSCCF model, the data items need to be correlated by some degree. If there is no correlation among data items, then our model would not work as it will filter out all uncorrelated data items.

VII. CONCLUSION

In this paper, we developed an innovative opinion prediction method in large scale cyber argumentation on multiple issues. Our method predicts how much a user would agree with a position on an issue based on the opinions of similar users on related issues. Our model achieved an excellent accuracy with a MAE value of 0.133 using collaborative filtering and correlations between positions across issues. We assessed the impact of number of positions predicted, and degree of correlation on the opinion prediction accuracy in multi-issue cyber argumentation. The method uses correlation to achieve high accuracy, thus it cannot work for discussions that are not related. Relevancy between discussions should be kept in mind when using this model. Using this model participants' opinions on related issues can be assessed even when they haven't explicitly discussed them. Additionally, discussions with a small number of participants can be analyzed more representatively. The predicted values can be used to impute missing values for different clustering algorithms for different opinionated group related analytical models. It can also be used to assess collective thoughts even when cyber argumentation on multiple issues is incomplete.

REFERENCES

[1] P. M. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games," Artificial Intelligence, vol. 77, no. 2, pp. 321–357, Sep. 1995.

[2] W. Kunz and H. W. J. Rittel, "Issues as elements of information systems," Inst. Urban and Regional Devt., Univ. Calif. at Berkeley, 1970.

[3] S. E. Toulmin. 1958. The Uses of Argument. Cambridge, UK: University Press, 1958.

[4] M. Klein, "The CATALYST Deliberation Analytics Server," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper, Nov. 2015.

[5] J. Sirrianni, X. Liu, and D. Adams, "Quantitative Modeling of Polarization in Online Intelligent Argumentation and Deliberation for Capturing Collective Intelligence," 2018 IEEE International Conference on Cognitive Computing (ICCC), pp. 57–64, 2018.

[6] S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing Value Imputation Based on Data Clustering," in Transactions on Computational Science I, M. L. Gavrilova and C. J. K. Tan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 128–138.

[7] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg, "Opinion Space: A Scalable Tool for Browsing Online Comments," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2010, pp. 1175–1184.

[8] T. Kriplean, J. Morgan, D. Freelon, A. Borning, and L. Bennett, "Supporting Reflective Public Thought with Considerit," in

 ISBN: 978-1-61208-757-3

Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, New York, NY, USA, 2012, pp. 265–274.

[9] M. Boireau, "Determining Political Stances from Twitter Timelines: The Belgian Parliament Case," in Proceedings of the 2014 Conference on Electronic Governance and Open Society: Challenges in Eurasia, New York, NY, USA, 2014, pp. 145–151.

[10] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, "#ISISisNotIslam or #DeportAllMuslims?: Predicting Unspoken Views," in Proceedings of the 8th ACM Conference on Web Science, New York, NY, USA, 2016, pp. 95–106.

[11] O. Fraisier, G. Cabanac, Y. Pitarch, R. Besançon, and M. Boughanem, "Stance Classification Through Proximity-based Community Detection," in Proceedings of the 29th on Hypertext and Social Media, New York, NY, USA, 2018, pp. 220–228.

[12] M. QIU, "Mining user viewpoints in online discussions," Dissertations and Theses Collection (Open Access), pp. 1–119, Jan. 2015.

[13] S. Gottipati, M. Qiu, L. Yang, F. Zhu, and J. Jiang, "Predicting User's Political Party Using Ideological Stances," in Social Informatics, 2013, pp. 177–191.

[14] M. Qiu, Y. Sim, N. A. Smith, and J. Jiang, "Modeling User Arguments, Interactions, and Attributes for Stance Prediction in Online Debate Forums," in Proceedings of the 2015 SIAM International Conference on Data Mining, 2015, pp. 855–863.

[15] X. Su and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Adv. in Artif. Intell., vol. 2009, pp. 4:2–4:2, Jan. 2009.

[16] B. Li, "Cross-Domain Collaborative Filtering: A Brief Survey," in 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011, pp. 1085–1086.

[17] P. Singh and G. J. Gordon, "Relational Learning via Collective Matrix Factorization," in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2008, pp. 650–658.

[18] X. Liu, E. Khudkhudia, L. Wen, and V. Sajja, "An Intelligent Computational Argumentation System for Supporting Collaborative Software Development Decision Making," Artificial Intelligence Applications for Improved Software Engineering Development: New Prospects, pp. 167–180, 2010.

[19] S. Sigman and X. F. Liu, "A computational argumentation methodology for capturing and analyzing design rationale arising from multiple perspectives," Information & Software Technology, vol. 45, pp. 113–122, 2003.

[20] X. (Frank) Liu, E. C. Barnes, and J. E. Savolainen, "Conflict Detection and Resolution for Product Line Design in a Collaborative Decision Making Environment," in Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, New York, NY, USA, 2012, pp. 1327–1336.

[21] M. R. Smith, M. S. Gashler, and T. Martinez, "A hybrid latent variable neural network model for item recommendation," in 2015 International Joint Conference on Neural Networks (IJCNN), 2015, pp. 1–7.

[22] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Matrix Factorization and Neighbor Based Algorithms for the Netflix Prize Problem," in Proceedings of the 2008 ACM Conference on Recommender Systems, New York, NY, USA, 2008, pp. 267–274.

[23] A. Mnih and R. R. Salakhutdinov, "Probabilistic Matrix Factorization," in Advances in Neural Information Processing Systems 20, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1257–1264.

[24] S. H. Schwartz, "An Overview of the Schwartz Theory of Basic Values" Online Readings in Psychology and Culture [Online]. Available from: https://scholarworks.gvsu.edu/orpc/vol2/iss1/11

 ISBN: 978-1-61208-757-3