TA-COS 2016
Text Analytics for Cybersecurity and Online Safety

23 May 2016

ABSTRACTS

Editors:

Guy De Pauw, Ben Verhoeven, Bart Desmet, Els Lefever

Workshop Programme

14:00 - 15:00 – Keynote

Anna Vartapetiance and Lee Gillam, *Protecting the Vulnerable: Detection and Prevention of Online Grooming*

15:00 - 16:00 – Workshop Papers I

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch and Derek Ruths, *A Web of Hate: Tackling Hateful Speech in Online Social Spaces*

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven and Walter Daelemans, *A Dictionary-based Approach to Racism Detection in Dutch Social Media*

16:00 - 16:30 – Coffee break

16:30 - 18:00 – Workshop Papers II

Adeola O Opesade, Mutawakilu A Tiamiyu and Tunde Adegbola, *Forensic Investigation of Linguistic Sources of Electronic Scam Mail: A Statistical Language Modelling Approach*

Richard Killam, Paul Cook, Natalia Stakhanova, *Android Malware Classification through Analysis of String Literals*

Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy and Norman Sadeh, *Demystifying Privacy Policies with Language Technologies: Progress and Challenges*

Workshop Organizers

Guy De Pauw	CLiPS - University of Antwerp, Belgium
Ben Verhoeven	CLiPS - University of Antwerp, Belgium
Bart Desmet	LT3 - Ghent University, Belgium
Els Lefever	LT3 - Ghent University, Belgium

Workshop Programme Committee

Walter Daelemans (chair)	CLiPS - University of Antwerp, Belgium
Veronique Hoste (chair)	LT3 - Ghent University, Belgium
Fabio Crestani	University of Lugano, Switserland
Maral Dadvar	Twente University, The Netherlands
Lee Gillam	University of Surrey, UK
Chris Emmery	University of Antwerp, Belgium
Giacomo Inches	Fincons Group AG, Switzerland
Eva Lievens	Ghent University, Belgium
Shervin Malmasi	Harvard Medical School, USA
Nick Pendar	Skytree Inc, USA
Karolien Poels	University of Antwerp, Belgium
Awais Rashid	Lancaster University, UK
Cynthia Van Hee	Ghent University, Belgium
Anna Vartapetiance	University of Surrey, UK

Preface

Text analytics technologies are being widely used as components in Big Data applications, allowing for the extraction of different types of information from large volumes of text. A growing number of research efforts is now investigating the applicability of these techniques for cybersecurity purposes. Many applications are using text analytics techniques to provide a safer online experience, by detecting unwanted content and behavior on the Internet. Other text analytics approaches attempt to detect illegal activity on online networks or monitor social media against the background of real-life threats. Alongside this quest, many ethical concerns arise, such as privacy issues and the potential abuse of such technology. The first workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016) aims to bring together researchers that have an active interest in the development and application of such tools.

Following our call for papers, we received papers on a wide range of topics and with the help of our varied team of reviewers were able to select the most relevant and most interesting contributions. The first two papers that are presented at this workshop deal with the issue of identifying hate speech on social media. Saleem and colleagues describe a technique that automatically detects hateful communities, while Tulkens *et al.* present research on how to develop techniques that idenfity hateful words and phrases on social media.

In the second session of this workshop Opesada *et al.* present a study on the origin of 419 Scam e-mails, using text classification techniques that identify varieties of English. Killam *et al.* identify malware on the Android platform by using text analytics on the apps' binary files. Finally, Wilson *et al.* describe work on developing techniques that can aid people in understanding the often lengthy and complex terms of use that they agree to online.

We are very pleased with this wide variety of topics of the submitted papers and are furthermore very pleased to be able to kick off our workshop with a keynote lecture by Anna Vartapetiance of the University of Surrey's Centre for Cyber Security. She will present ongoing research on the automatic detection of online grooming. We are sure that the presentations at TA-COS 2016 will trigger fruitful discussions and will help foster the awareness of the increasingly important role text analytics can play in cybersecurity applications.

The TA-COS 2016 Organizers,
Guy De Pauw
Ben Verhoeven
Bart Desmet
Els Lefever
`www.ta-cos.org`

Keynote

23 May, 14:00 – 15:00
Chairperson: Guy De Pauw

Protecting the Vulnerable: Detection and Prevention of Online Grooming

Anna Vartapetiance and Lee Gillam

The 2012 EU Kids Online report revealed that 30% of 9-16 year-olds have made contact online with someone they did not know offline, and 9% have gone to an offline meeting with someone they first met online. The report suggests that this is "rarely harmful", but is hoping against harm really the wisest course of action?
This talk presents details of our ongoing research and development on the prevention and detection of unsavoury activities which involve luring vulnerable people into ongoing abusive relationships. We will focus specifically on online grooming of children, discussing the potential to detect and prevent such grooming, and relevant theories and systems.
The talk will address some of the challenges involved with the practical implementation and use of such safeguards, in particular with respect to legal and ethical issues. We conclude by discussing the opportunities for protecting further groups vulnerable to grooming for emotional, financial, or other purposes.

Biography
Dr Anna Vartapetiance, is a graduate entrepreneur and postdoctoral researcher at the University of Surrey's Centre for Cyber Security, as well as a committee member for the BCS ICT Ethics Specialist Group. Anna is currently on the advisory committee of the "Automatic Monitoring for Cyberspace Applications (AMiCA)" project as an international expert, and an associate of the Internet Service Providers Association (ISPA). She is also a member of the International Federation of Information Processing (IFIP) Special Interest Group 9.2.2 Framework on Ethics of Computing (SIG 9.2.2) and the Working Group on Social Accountability and Computing (WG 9.2).
Prior to her work as entrepreneur and postdoctoral researcher, she was awarded a PhD from the Department of Computer Science at the University of Surrey for her work on *deception detection* using Natural Language Processing (NLP) to develop (semi-)automated detection systems. Her research has found application in systems that enhance outcomes and issues related to national DNA databases, online gambling, virtual worlds and machine ethics and has been published in over 15 peer reviewed journal papers, proceedings and book chapters.
Currently, Anna is working on *parental controls* with a Child Online Safety project prototype for the detection / prevention of online grooming.

Workshop Papers I

23 May, 15:00 – 16:00
Chairperson: Véronique Hoste

A Web of Hate: Tackling Hateful Speech in Online Social Spaces

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch and Derek Ruths

Online social platforms are beset with hateful speech - content that expresses hatred for a person or group of people. Such content can frighten, intimidate, or silence platform users, and some of it can inspire other users to commit violence. Despite widespread recognition of the problems posed by such content, reliable solutions even for detecting hateful speech are lacking. In the present work, we establish why keyword-based methods are insufficient for detection. We then propose an approach to detecting hateful speech that uses content produced by self-identifying hateful communities as training data. Our approach bypasses the expensive annotation process often required to train keyword systems and performs well across several established platforms, making substantial improvements over current state-of-the-art approaches.

A Dictionary-based Approach to Racism Detection in Dutch Social Media

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven and Walter Daelemans

We present a dictionary-based approach to racism detection in Dutch social media comments, which were retrieved from two public Belgian social media sites likely to attract racist reactions. These comments were labeled as racist or non-racist by multiple annotators. For our approach, three discourse dictionaries were created: first, we created a dictionary by retrieving possibly racist and more neutral terms from the training data, and then augmenting these with more general words to remove some bias. A second dictionary was created through automatic expansion using a `word2vec` model trained on a large corpus of general Dutch text. Finally, a third dictionary was created by manually filtering out incorrect expansions. We trained multiple Support Vector Machines, using the distribution of words over the different categories in the dictionaries as features. The best-performing model used the manually cleaned dictionary and obtained an F-score of 0.46 for the racist class on a test set consisting of unseen Dutch comments, retrieved from the same sites used for the training set. The automated expansion of the dictionary only slightly boosted the model's performance, and this increase in performance was not statistically significant. The fact that the coverage of the expanded dictionaries did increase indicates that the words that were automatically added did occur in the corpus, but were not able to meaningfully impact performance. The dictionaries, code, and the procedure for requesting the corpus are available at: `https://github.com/clips/hades`.

Workshop Papers II

23 May, 16:00 – 16:30
Chairperson: Walter Daelemans

Forensic Investigation of Linguistic Sources of Electronic Scam Mail: A Statistical Language Modelling Approach

Adeola O Opesade, Mutawakilu A Tiamiyu, Tunde Adegbola

Electronic handling of information is one of the defining technologies of the digital age. These same technologies have been exploited by unethical hands in what is now known as cybercrime. Cybercrime is of different types but of importance to the present study is the 419 Scam because it is generally (yet controversially) linked with a particular country - Nigeria. Previous research that attempted to unravel the controversy applied the Internet Protocol address tracing technique. The present study applied the statistical language modelling technique to investigate the propensity of Nigeria's involvement in authoring these fraudulent mails. Using a hierarchical modelling approach proposed in the study, 28.85% of anonymous electronic scam mails were classified as being from Nigeria among four other countries. The study concluded that linguistic cues have potentials of being used for investigating transnational digital breaches and that electronic scam mail problem cannot be pinned down to Nigeria as believed generally, though Nigeria could be one of the countries that are prominent in authoring such mails.

Android Malware Classification through Analysis of String Literals

Richard Killam, Paul Cook and Natalia Stakhanova

As the popularity of the Android platform grows, the number of malicious apps targeting this platform grows along with it. Accordingly, as the number of malicious apps increases, so too does the need for an automated system which can effectively detect and classify these apps and their families. This paper presents a new system for classifying malware by leveraging the text strings present in an app's binary files. This approach was tested using over 5,000 apps from 14 different malware families and was able to classify samples with over 99% accuracy while maintaining a false positive rate of 2.0%.

Demystifying Privacy Policies with Language Technologies: Progress and Challenges

Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy and Norman Sadeh

Privacy policies written in natural language are the predominant method that operators of websites and online services use to communicate privacy practices to their users. However, these documents are infrequently read by Internet users, due in part to the length and complexity of the text. These factors also inhibit the efforts of regulators to assess privacy practices or to enforce standards. One proposed approach to improving the status quo is to use a combination of methods from crowdsourcing, natural language processing, and machine learning to extract details from privacy policies and present them in an understandable fashion. We sketch out this vision and describe our ongoing work to bring it to fruition. Further, we discuss challenges associated with bridging the gap between the contents of privacy policy text and website users' abilities to understand those policies. These challenges are motivated by the rich interconnectedness of the problems as well as the broader impact of helping Internet users understand their privacy choices. They could also provide a basis for competitions that use the annotated corpus introduced in this paper.