Arcada Working Papers 8/2016
ISSN 2342-3064
ISBN 978-952-5260-75-5

Working Papers Presented in Arcada Workshop on Analytics in June 8, 2016

Göran Pulkkis[i] (Ed.)

Abstract

Department of Business Management and Analytics in Arcada University of Applied Sciences arranged a Workshop on Analytics in June 8, 2016. Five Working Papers presented in this workshop are published in this report.

CONTENTS

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [goran.pulkkis@arcada.fi]

Online Trust in the European Media Context

Hilma Immonen[i]

Abstract

This working paper analyses the concept and state of online trust, the confidence formed towards Social Networking Sites (SNS). SNS are used frequently for different activities that used to take place mainly offline. According to the recently published Eurobarometer 84.3 survey data, trust towards SNS is low despite of increasing usage rates. SNS are increasingly used as news source in European countries and this paper aims to identify what affects individuals to form trust towards online social networks. Statistical testing indicates that online trust is in most European countries formed through a more intense use of SNS and also the accountability of traditional media is in politically less stable countries associated with individual online trust. Even though online trust emerges from a frequent use of SNS, our analysis shows that social media use is rather well-grounded than imprudent among the majority of Europeans.

Keywords: online social networking, online trust, media trust

1 INTRODUCTION

Online activities extend from producing online content like text or videos to organizing the exchange of physical goods or services on online platforms. As former offline activities like discussion and sales take increasingly place online, the relevance of online social networks keeps increasing. Social Networking Sites (SNS) offer for some an additional, for some an alternative channel for civic and political activism (Schlozman et al. 2010; Foos et al. 2015), take an important role in news and information consumption (Norris 2001) and fasten and facilitate peer-to-peer communication for private means. In the last years SNS have received even more attention as not only companies have moved to the web constructing an Internet of Things, but also the platform economy is replacing and supplementing traditional trade and creation of value also within the

[i] Arcada University of Applied Sciences, Finland, Department of Culture and Communication, [immonenh@arcada.fi]

boundaries of SNS (Choudary et al. 2016). SNS steadily continue to grow as an additional public sphere, where offline activities are monitored and transferred to.

However simultaneously SNS appear to cast doubt among Internet users (European 2014) and in the European context, the confidence towards online social networking is by far more intense in some countries than in others. Macro-level observations imply, that in countries where traditional media like the radio and press are held more accountable like in Sweden or Finland, fewer people trust social media, despite high usage rates. On the other hand Eastern and Southern European national media systems suffer of lower trust and the confidence formed towards SNS is comparably higher than in other parts of Europe. These empirical observations suggest that not trustworthy perceived media content motivates individuals to trust SNS to a higher extent. In the last years especially security concerns due to the trade with personal information data have harmed consumer confidence towards Facebook and other social networking sites (Debatin et al. 2009). Still, despite all public criticism and data scandals, SNS keep attracting users from all over the world. Some scholars argue that the need to belong is often stronger than existing security concerns (Debatin et al. 2009). Also network effects have been discovered in the last years on the web market implying a fast overtake by a new visionary application or webpage that leads to the downfall of others (Evans & Schmalensee 2016). This paper evaluates the formation of online trust through the intensity of usage and the novelty compared to traditional media.

The lack of trust towards social media suggests that people are not willing to take risks in online environments and this paper aims to evaluate online trust in the European context and explain how online trust is associated with the individual habits of social media use and trust towards traditional media. Do attitudes towards traditional media influence the generation of online trust? Then, as the intentions and motivation behind Online Social Networking (OSN) can vary, does online trust depend from the manner of online behavior? Finally, does the extent of OSN use predict higher online trust or are security disputes ignored by users even though they do not trust SNS? At best the results of my analysis can lead to conclusions about the strength of network effects in the European media context and identifying differences, which lead to a more critical online engagement in some countries. This paper is part of a research project about Sharing Economy[i], which is an urban form of civic participation aimed at effective sharing and common use of different resources like vehicles and groceries, enabled and coordinated online, also through online social networks (Nylund 2015).

I will start by reviewing the novelty features that SNS have brought to media use. Then the concept of online trust will be reviewed and possible influences of the use of SNS and the perception on traditional media institutions on the generated trust will be outlined. This article concludes that online trust is in some cases predictable by trust towards traditional media institutions however if so, then associated positively with how traditional media institutions are perceived. Further in some European states consuming news from SNS is significantly associated to online trust and the extent of social media use also is associated with trust towards these online networks.

[i] Sharing Economy: Peer-to-peer based sharing of access to goods and services

2 THE RISE OF SNS

SNS change reporting and behaving in reality compared to the TV, press or radio as a mirror of the society. Online networks act as platform that enable, fasten and facilitate different online and offline activities like discussion and sales which take place among individual users of the platforms. The understanding of media forming a prism of reality, being an integral, institutional and affecting part of the society the passive audience either accepts or does not accept the given information (Davis 2001), cannot be adopted to SNS because of the interactive and communicative character of online behavior. Davis' comprehension rather suits the traditional media presentation of the reality. Online social networks function as a meeting point where everyone can report about and contribute to the creation of reality. Lietsala & Sirkkunen (2008) define five characteristics that comprise social media: „A space to share content; participants create, share or evaluate all or most of the content themselves; It is based on social interaction; all content has an URL to link it to the external networks; all actively participating members of the site have their own profile page to link to other people, to the content, to the platform itself and to the possible application".

2.1 The Emerge of Online Trust

According to Nissenbaum (2001): ''Trust is an extraordinarily rich concept, covering a variety of relationships, conjoining a variety of objects. One can trust (or distrust) persons, institutions, governments, information, deities, physical things, systems, and more''. Trust can be formed in the online environment towards online actors, online activities or procedures and the quality of online content, or simultaneously all of them as they constitute social media as an institution. Trust and distrust are formed by earlier experience with the trusted object and guide the willingness to take a future risk. Like social capital is created offline through trust towards individuals in one's environment and is e-capital closely linked to online trust. E-capital is according to Merisalo et al. (2008) "the possibility, ability and willingness to use ICT, electronic services and social media, resulting in benefits to the users, the economy and society". The relationship of trust and e-capital is reciprocal, as successfully experienced online engagement leads to the formation of trust and trust in turn motivates individuals for further online behavior. The formation of online trust is important for online civic engagement and commercial online behavior for instance but also for SNS being used for personal communication and experience sharing.

Just as in real-life where weak and strong relationships (ties) define the strength of social capital (Putnam 2000), also on SNS one's peer groups take an important role as news, videos and pictures posted by other users are vital for SNS to exist. On the mostly used SNS Facebook (Statista 2016), individuals usually become friends with already familiar offline peers which also remain more trusted than loose internet acquaintances (Ellison et al. 2007). Content that is produced or reproduced by one's closer peers is easier and faster accessible than those of strangers or not-befriended users. The concept of trust is often linked to social capital, which as Fukuyama puts it is „a capability that arises from the prevalence of trust in a society or in certain parts of it'' (Fukuyama 1995:26) and the role of one's online peers needs to be taken into account when as-

sessing trust formed towards SNS. Social capital describes the individual's willingness to interact with others and is highly influenced by trust and positive attitudes that are formed towards others.

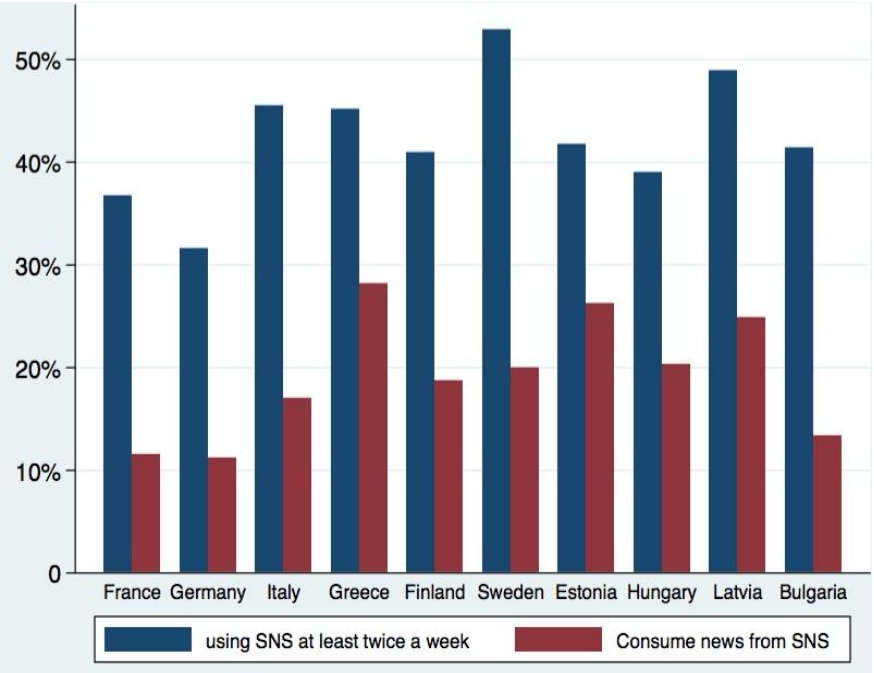
Most research on online trust has so far concentrated on its role for e-commerce. The formation of trust towards a certain online product, app or payment transaction is easier to measure as the abstraction level is not as high as when asking individuals about trust towards the institution of SNS, which contains millions of users, organizations and content. Some scholars also make a difference between user trust arising from interactions with other users and system trust explaining one's fulfilled expectation of using a device or system (Sherchan et al. 2013). Wang & Emurian (2005) have identified distrust in e-commerce arising mainly from concerns about security and anonymity of personal data. According to many scholar also general attitudes towards technological developments and the overwhelming novelty of communicating online has formed distrust towards SNS and the internet (Nissenbaum 2004; Wang & Emurian 2005). Serious security concerns should also lead to a decline of social media use for non-commercial purposes. However the opposite seems to be the case as more and more individuals decide to connect online with their peers and user-driven advantages apparently outweigh systemic problems.

2.2 The Choice of Media

Only a few studies so far have investigated the associations existing between traditional and new media use, and seldom distrust towards national media institutions is considered driving individuals to look for alternative sources online. However media use of individuals is driven by the offered content and if traditional media fails in producing credible news that individuals rate as trustworthy consuming again, individuals have in times of web 2.0 several channels where to search for news information instead. According to an up-to-date study by Reuters the majority of European youth prefers to consume news from SNS rather than watch news on TV (Digital News Report 2016). Consumption of SNS in some European counties is shown in Figure 1.

SNS constantly succeed to create new ways in attracting users when simultaneously TV and press are loosing audience. Also the Eurobarometer reveals that the trust towards press, radio, and TV are especially among younger generations increasingly distrusted (European 2014). Distrust is formed especially towards media institutions where political actors are involved that are neither trusted by the consumers. In Italy for example, long-term Prime Minister Berlusconi owned the largest network of private television channels, and with increasing distrust he perceived over the course of several scandals during his career, the critical audience turned to other channels (Durante & Knight 2012). In Bulgaria bribing journalists and promoting them on political grounds is hard to prove but widely perceived by the citizens and gives also a reason to look for alternative information channels (Sipos 2014). In Bulgaria and Italy for instance SNS perceive higher trust than press which is according to the Freedom House Index only partly free (Freedom 2016). Corruption and biased reporting can occur also both in private and in state-owned media. Tsafi & Ariely (2013) have found in their cross-sectional analysis of 44 countries that state-owned media is rather perceived as trustworthy by citizens, if the

political system responds in a more democratic way to citizen demands. Trust towards TV, radio, written press and SNS in different European nations is shown in Figure 2.



. *Figure 1. Consumption of SNS*

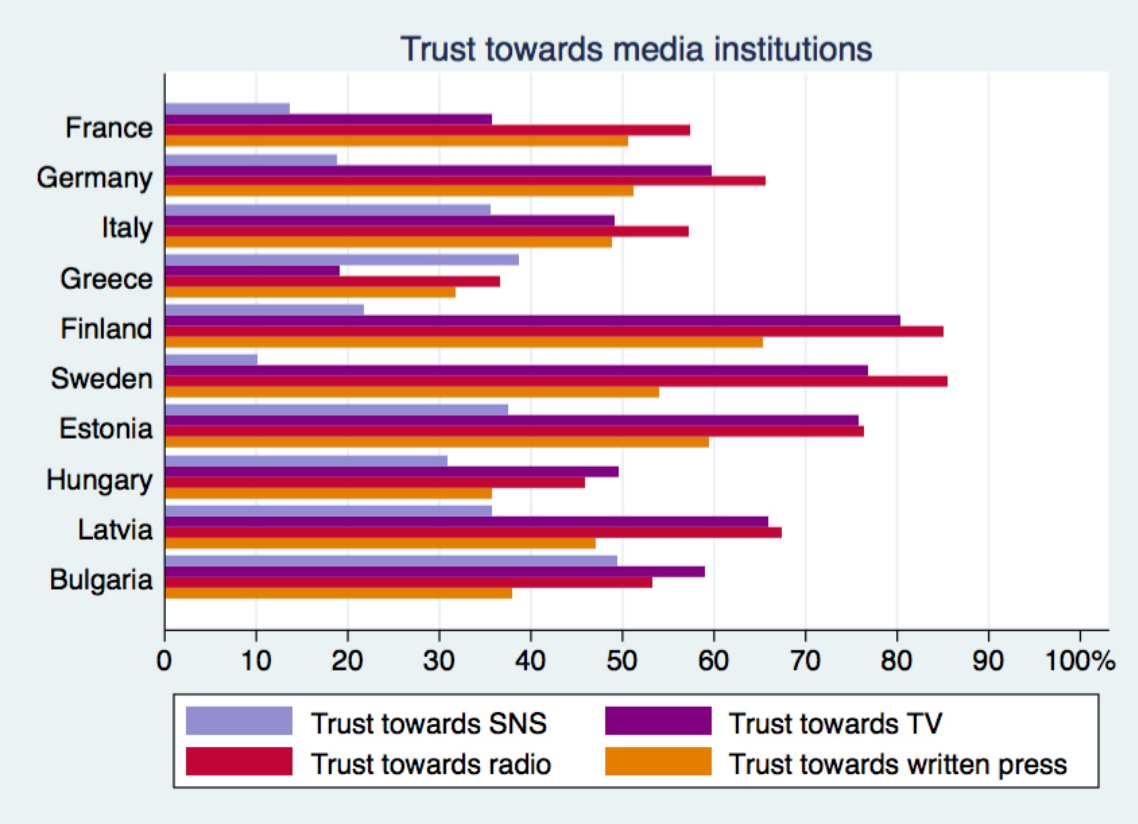


Figure 2. Trust towards TV, radio, written press and SNS in different European nations

In the European context media systems are often funded, controlled or owned partly or fully by the state, so media trust is linked to political power structures. The state-owned channels in Finland or public-service broadcasting like in Germany, savor of high trust from the audience and are held accountable according to the figures of the Eurobarometer. In Finland and Sweden social media is in intense use, but it is not common to consume news from SNS. Tsafi & Ariely (2013) have found in their cross-sectional analysis of 44 countries that Internet use is associated with distrust towards TV, press and radio.

Biased reporting is claimed to be the main predictor of distrust towards media. Media institutions interacting with opposed interests are not likely being consumed, trust is mainly formed according to one's own group identification (Gunther 1992). The consequence for a distrusting media audience is to look for alternative channels to get informed. On SNS where one can get in touch easier with users having similar attitudes and interest being it one's close friends or also strangers, the choice of ones interaction partners is driven by the same group identification that Gunther (1992) talks about. The more frequent use of SNS and the constant interaction with one's peer groups sharing content, creates social capital and trust, as basically the interaction is merely an image of how interaction used to take place offline. SNS offer a more consistent environment according to one's attitudes that motivates to spread and consume news and that's why using SNS not only creates trust towards users, content and the networks, but also in turn leads to higher use, when earlier experience is perceived as successful. From these theoretical consideration two hypotheses for the further analysis are derived:

H1: A more frequent and intense individual use of SNS, is associated with a higher trust towards SNS.

H2: The lower trust towards traditional media institutions is, the higher is the trust build towards online social networks.

3 DATA AND METHODS

In order to test the theoretical assumptions, the Eurobarometer 84.3 data set is used.The data was collected in November 2014 and published in May 2016. All in all the dataset contained 32.833 observations (respondents) and 835 variables. To my knowledge this is the first freely accessible survey data containing variables about both social media use and online trust in all countries of the EU. (Standard 2016)

The binary dependent variable online trust has two answering categories, whether the respondent trusts or distrusts online social networks. The independent variables for the first hypothesis contain the questions about one's OSN frequency with recoded values: Never as reference category, Sometimes and daily use. Further the individual intensity of social media use can be captured with the respondent mentioning SNS to be the first source for consuming national news from a list of different news sources. The trust towards traditional media institutions is measured by binary variables indicating either trust or distrust towards the institutions TV, written press and radio. I also control for gender and age, as youth is claimed to use SNS more and according to a current study by Reuters, also women are more prone to using SNS than men.

For the testing of my theoretical assumptions on online trust, 10 country cases with different media background were chosen. Estonia, Latvia, Finland, Sweden, Germany, France, Italy, Greece, Bulgaria and Hungary are analyzed. With these cases conclusions about the formation of online trust in different politically influenced media context can be made, also these cases show on macro level analysis differences in social media use in general and for news consumption.

4 ANALYSIS

The statistical analysis shows significant results for some of the assumed relationships. Online trust is dependent at least in some political environments from trust towards other media and also the time dedicated to SNS as well as trust formed by individuals towards online news content. According to the Nagelkerke's R2 estimation, which predicts the likelihood of the full model perfectly predicting the emerge of online trust, we can see that the proposed variables best predict online trust in Bulgaria (45% of outcome predicted), Italy (37%), Hungary (28%) and Greece (27,9%) – see Table 1. However in case of the other countries, the individual level analysis only predicts correctly the outcome of around 20% (Latvia, Estonia and Finland) or even less in Germany, Sweden and France. From this one must conclude, that in the Central European, Nordic and Baltic countries some other factors, which were not included in this regression analysis contribute to a higher extent to the formation of trust towards SNS.

Table 1. The predictors of online trust

	Greece	Italy	France	Germany	Finland	Sweden	Estland	Lithuania	Bulgaria	Hungary
Not trusting political online content	**0.641*** ** (3.64)	**0.563*** (-2.41)	**0.951*** ** (3.64)	0.389 (1.51)	**1.578*** ** (6.35)	**1.271*** ** (4.37)	0.449 (1.67)	0.365 (1.50)	**1.216*** ** (4.78)	0.171 (0.82)
TV trust	**1.016*** ** (3.58)	**1.237*** ** (4.49)	**0.609*** (1.98)	0.697 (1.75)	0.547 (1.21)	0.347 (0.67)	0.774 (1.59)	0.284 (0.92)	**1.288*** ** (3.42)	0.406 (1.60)
Radio trust	**0.596*** (2.54)	**0.971*** ** (3.45)	-0.229 (-0.63)	-0.363 (-0.83)	-0.342 (-0.72)	-0.760 (-1.36)	0.465 (0.89)	0.367 (1.06)	0.288 (0.70)	**0.596*** (2.30)
Written press trust	**0.573*** (2.36)	0.335 (1.23)	-0.228 (-0.67)	**0.960** ** (2.70)	**0.918** ** (2.63)	0.589 (1.68)	0.584 (1.59)	**0.903** ** (2.84)	**0.980** ** (2.60)	**1.360*** ** (5.52)
Online Social Networking:										
Never	Reference									
Sometimes	1.105*** (4.10)	0.912* (2.39)	0.593 (1.47)	**0.902** ** (2.59)	-0.393 (-1.12)	0.600 (1.04)	0.638 (1.25)	1.919*** (3.21)	**1.188** ** (3.21)	**0.645*** (2.21)
(Almost everyday)	**1.438*** ** (5.45)	**1.776*** ** (4.51)	**1.032** ** (2.90)	**1.227** ** (3.27)	0.230 (0.68)	**1.513** ** (2.86)	**1.049*** (2.17)	**1.919*** ** (5.28)	**1.798*** ** (4.59)	**0.980** ** (3.24)
Newssource	0.273 (1.38)	**1.132*** ** (4.12)	**0.608*** (2.02)	0.257 (0.82)	0.394 (1.24)	0.482 (1.43)	0.0323 (0.12)	-0.454 (-1.71)	**0.555** (1.72)	**0.918*** ** (3.87)
Age	-0.00587 (-0.89)	0.00526 (0.61)	-0.00191 (-0.22)	0.000234 (0.03)	0.00790 (0.91)	0.0133 (1.35)	**0.0215*** (2.39)	**0.0221** ** (2.63)	-0.0181 (-1.63)	-0.000552 (-0.08)
Gender	0.0665 (0.38)	-0.305 (-1.33)	-0.307 (-1.19)	**0.510*** (2.12)	-0.00520 (-0.02)	**-0.720*** (-2.36)	-0.167 (-0.63)	-0.385 (-1.66)	0.0471 (0.19)	-0.0666 (-0.32)
Constant	-2.042*** (-4.53)	-3.089*** (-4.84)	-2.549*** (-4.30)	-3.413*** (-5.57)	-3.181*** (-4.31)	-4.310*** (-4.75)	-3.810*** (-5.07)	-3.385*** (-5.50)	-2.409*** (-3.50)	-2.634*** (-5.32)
N	697	459	540	518	511	673	314	406	416	561
Nagelkerke's R2	0.279	0.376	0.132	0.148	0.189	0.142	0.180	0.224	0.454	0.289

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Beginning with the demographics, the logistic regression model shows only for Sweden and Germany effects in the respondent's gender. In Sweden females are slightly less likely to trust SNS than males and in Germany males perceive SNS as more trustworthy. Also age is only in a few countries associated with online trust: In Latvia and Estonia the trust towards SNS increases slightly with the age of the respondent. This indicates that in the Baltic countries where being connected online is very common, online activities are chosen carefully.

A more frequent use of SNS is significantly and positively associated to an individuals trust towards SNS in all analyzed countries. The only country where this finding does not apply is Finland. In Sweden diligent online users are even 4 times more likely to trust SNS compared to the reference category which consisted of non-users. Also the type of online activity has in some countries a significant effect on online trust. In Italy, France and Hungary SNS are more trusted when they are also used as main news source. In the other countries analyzed, online trust is generated independently from whether SNS are the main news source. In conclusion, trust towards SNS is not essentially influenced by the type of individual online activity but rather from the time devoted all together to online participation. This conclusion is further supported by the finding, that those not trusting political online content, are significantly predicted still to more likely trust SNS in half of the analyzed countries. Online trust is formed independently from the generation of trust towards political opinion postings for instance. Only in Italy where even two thirds of all online users also consume news mainly from SNS, is distrust with political online content also associated with distrust towards SNS in general.

When turning to the relationship of traditional media trust and online trust, it seems that the patterns holding on macro-level do not allow to make conclusions on the individual level and to assume that higher online trust arises of no trust towards traditional news media. In fact, the trust formed towards TV and radio is especially in countries that suffer of low media trust, positively and significantly associated to online trust. Italians and Greeks that trust in TV and radio are more likely to trust SNS, further also in France and Bulgaria those trusting TV are also more likely trusting SNS. Also in Hungary radio trust is positively associated to online trust. This implies that media is in many countries either trusted or not trusted. Even though the association between radio and online trust is significant in Bulgaria, Greece and Hungary, it is fairly weak. For all other countries no significant relationships between radio/TV and online trust occur in this analysis. Media trust is most likely to effect online trust in countries, where traditional media is less trusted and SNS are trusted by comparably wider population.

The effect of trust towards the written press in turn, shows consistencies between European countries: In Greece, Germany, Finland, Latvia, Bulgaria and Hungary, those who trust the national written press, also are more likely to trust SNS. This can be explained by the frequent and easy reposting of online articles in the online environment on SNS. Either the written press benefits in terms of trust from SNS as online articles are distributed among social networks, or SNS benefit from a trusted news content that is easily distributed to one's online friends. However a significant amount of individuals shares trust both towards press and SNS.

Generally these findings indicate that more frequently connected users trust SNS in general to a higher extent. Even if SNS are trusted as an institution where one can connect with friends to an increasing extent all over Europe, not all content is trusted. Those who perceive political content as trustworthy are not necessarily perceiving SNS being trustworthy. The analysis mainly proves the relationship between increases in online activities and trust. Even though SNS take an increasingly large role in news consumption, Europeans do not not trust this information blind. The majority of Europeans still perceives SNS with doubt and consumes news on an informed and critical basis. And this analysis provide evidence that there is a substantial part of European population, especially in Southern and Eastern European countries, that trust whether traditional nor new media.

5 CONCLUSIONS

SNS are in general used frequently and increasingly all over Europe and despite this development a majority of the users does not trust SNS. This paper aimed at identifying SNS as a competing channel to traditional media through its increase as news source and whether the accountability of SNS can be expected to increase if traditional media is increasingly less trusted. The analysis shows that unsurprisingly online trust is associated to a more frequent use of SNS. However the analysis presents only limited evidence for the assumption that when news are consumed online, SNS are in general perceived more trustworthy. In fact distrust towards political online content was in some country samples significantly associated with online trust, which means that other factors are more powerful in predicting how online trust is formed. Consuming online news does not increase the likelihood of trusting SNS. These findings indicate that social media use is among the majority of Europeans well-grounded and usually it is not the information consumed online that creates online trust.

The Eurobarometer data set bares some limitations in investigating online trust. For testing the intensity of respondents SNS use, the data set does not contain other variables, which describe an individual's exact activities on SNS apart from news consumption. Unfortunately the data set does not contain variables about the respondents' perception of online security either. According to many scholars, this influences online trust and could perhaps rise the explanation power of the regression models. In future studies security concerns towards systemic faults of SNS and personal data concerns could be included in such an analysis as individuals may limit any kind of online behavior due to serious issues that might rise with the network platform.

The main purpose of using SNS is being connected with other individuals that are likely to be similar to the users. Using SNS as news source has increased due to the possibility of linking online articles to one's peers and actually most who trust SNS also trust the written press. As neither traditional media trust nor online news consumption showed consistent influence on online trust through all cases, this analysis provides evidence, that online trust is formed in different political and medial contexts out of different grounds. In Bulgaria, Hungary, Italy and Greece, where media in general is less trusted than in other parts of Europe being it on grounds of a state crisis or corruption, online trust among individuals is more likely being coupled to trust towards traditional media institutions, which in turn is often influences in some countries by media ownership or

political conditions like the degree of democracy (Tsfati & Ariely 2013). Online social networks are often trusted for the same reasons as TV, the written press or radio especially in those countries, where press is not rated as fully free, online affinity can be concluded to grow out of general media enthusiasm.

Finland is an exceptional case in this analysis as more frequent online use is not associated with online trust. A larger, more representative sample could provide a different result, or Finns can be concluded to be even more critical and less trusting towards SNS than other Europeans. The reasons for distrust towards SNS in Finland and also among the majority of Europeans could lay also in recent scandals that have shaken the accountability of SNS. Especially the most popular SNS Facebook keeps returning to the public discourse in critical headlines, last being accused of a biased presentation of news entries in the news feed, which appear on basis of human assessment on the worldwide most discussed topics (BBC 2016). Basically sorting news according to their popularity is nothing different from how newsworthy topics are chosen for news broadcast or newspapers, but Facebook is reported to have benefitting conservatively reported news. As press freedom effects which articles are being published and in 2016 still the majority of countries worldwide lack of freedom in written press, the selection of newsworthy articles can lead to rejecting some political views or critical articles about SNS themselves.

REFERENCES

Choudary, S. P., Van Alstyne, M. W., & Parker, G. G. 2016, *Platform Revolution How the networked Markets Are Transforming the Economy and How to Make Them Work for You*. New Yourk: W. W. Norton & Company, Inc., ISBN: 978-0-393-24913-2

Davis, R, 2001, *The Press and American Politics*: *The New Mediator*. Third Edition. Prentice Hall.

Debatin, B., Lovejoy, J. Horn, A.-K. & Hughes, B. 2009, Facebook and online privacy: Attitudes, behaviors, and unintended consequences, in: *Journal of Computer-Mediated Communication*, Vol 15, No. 1, pp. 83-108.

Digital News Report. 2016. Reuters Institute for the Study of Journalism. Accessed 1.6.2016. Published 2016. http://www.digitalnewsreport.org

Durante, R. & Knight, B. 2012, Partisan control, media bias, and viewer responses: evidence from Berlusconi's Italy, in: *Journal of the European Economic Association*, Vol. 10, No. 3, pp. 451-481.

Ellison, N., Steinfeld, C., & Lampe, C. 2007, The Benefits of Facebook „Friends“:Social Capital and College Students' Use of Online Social Network Sites, in: *Journal of Computer-Mediated Communication*, Vol. 12, No 4, pp. 1143-1168.

European Commission. 2014, Media Use in the European Union. Report. Accessed 25.5.2016. Published 2014. http://ec.europa.eu/public_opinion/archives/eb/eb82/eb82_media_en.pdf

Evans, D. & Schmalensee, R. 2016. Why Winner-Takes-All Thinking Doesn't Apply to the Platform Economy. Harvard Business Review. Accessed 9.5.2016. Published 2016. https://hbr.org/2016/05/why-winner-takes-all-thinking-doesnt-apply-to-silicon-valley?cm_sp=Article-_-Links-_-Comment

Foos, F. Kostadinov, L., Marinov, N., & Schimmelfennig, F.. 2015. Does Social Media Promote Civic Activism? Evidence from a Field Experiment. Accessed 6.6.2016. Published 2015. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2704356

Freedom House. Freedom of the Press 2016. Accessed 4.6.2016. Published 2016. https://freedomhouse.org/report/freedom-press/freedom-press-2016

Fukuyama, F. 1995, *Trust, the Social Virtues and the Creation of Prosperity*. New York: Free Press.

Gunther, A. C. 1992. Biased press or biased public? Attitudes toward media coverage of socialgroups, in: *Public Opinion Quarterly,* Vol. 56, No 2, pp. 147-167.

Lietsala, K. & Sirkkunen, E. 2008. Social Media: Introduction to the Tools and Processes of Participatory Economy. Hypermedia Laboratory Net Series, University of Tampere. Accessed 2.6.2016. Published 2008. https://tampub.uta.fi/bitstream/handle/10024/65560/978-951-44-7320-3.pdf?sequence=1

Merisalo, M., Makkonen, T., & Inkinen, T. 2008. Creative and Knowledge-Intensive Teleworkers' Relation to e-Capital in the Helsinki Metropolitan Area, in: *International Journal of Knowledge-Based Development (IJKBD).* Vol. 4. No 3.

Nissenbaum, H. 2001. Securing Trust Online: Wisdom or Oxymoron, in: *Boston University Law Review* Vol. 81, No 3, pp. 635-664.

Nissenbaum, H. 2004. Will Security Enhance Trust Online, or Supplant It? in: Kramer, R. & Cook, K. (Eds.) *Trust and Distrust in Organizations: Dilemmas and Approaches.* New York: Russel Sage Foundation.

Norris, P. 2001. *Digital divide.* Cambridge: Cambridge University Press.

Nylund, M. 2015. Jakamistalous: Urbaanin osallistumisen ja kansalaisaktivismin uusi muoto. Arcada Working Papers. No 4/2015.

Putnam, R. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster.

Schlozman, K., Verba, S., & Brady, H. 2010. Weapon of the strong? Participatory inequality and the Internet, in: *Perspectives on Politics.* Vol. 8, No. 2, pp. 487-509.

Sherchan, W., Surya, N., & Paris, C. 2013. A Survey of Trust in Social Networks, in: *ACM Computing Surveys.* Vol. 45, No. 4.

BBC News. 2016. Social media 'outstrips TV' as news source for young people. Accessed 20.6.2016. Published 15.6.2016. http://www.bbc.com/news/uk-36528256

Sipos, Z. 2014, Bulgaria: Murky ownership, censorship and corruption in the media. Accessed 2.6.2016. Published 22.12.2014. https://www.indexoncensorship.org/2014/12/bulgaria-murky-ownership-censorship-and-corruption-in-the-media/.

Standard Eurobarometer 84, 2016. Accessed 6.6.2016. Published 2016. https://data.europa.eu/euodp/en/data/dataset/S2098_84_3_STD84_ENG

Statista, 2016. The Statistics Portal Accessed 6.6.2016. Published 2016. http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Tsafi, Y. & Ariely, G. 2013. Individual and Contextual Correlates of Trust in Media across 44 Countries, in: *Communication Research*. May 2, pp. 1-23.

Wang, Y. D. & Emurian, H. 2005. An Overview of Online Trust: Concepts, Elements, and Implications, in: *Computers in Human Behavior.* Vol. 21, No 1, pp. 105-125.

.

Reflections on Classification Models for Detecting Hate and Violence*

Shuhua Liu[i]

Abstract

Automatic detection of hate and violence content on the web is essential for user's online wellbeing and parental control solutions. Web classification models in the early days are limited by the methods and data available. In our research we revisited the problem of web content classification with new methods and techniques including elements of topic extraction, content similarity analysis, sentiment analysis and LDA topic modelling. In this paper we review recent studies on hate and violence content detection, and report our experience and learnt lessons with different classification methods and models.

Keywords: content classification models, detecting hate and violence, topic modelling, content similarity, sentiment analysis, online safety solutions

1 INTRODUCTION AND BACKGROUND

The Internet has become our primary means of communication and information resource with unprecedented global reach. We sometimes forget that the web hosts both "the best and the worst written products of humanity" (Levmore & Nussbaum 2011).

"Being open for use and abuse, the Internet is therefore saturated with content that would unlikely be entertained by conventional media. As it provides cheap, instantaneous, and decentralized distribution, numerous points of access, no necessary ties to geography, no simple system to identify content, as well as sophisticated encryption tools, the Internet has become an asset for hate groups to transmit propaganda and provide information about their aims, allow an exchange between like-minded individuals, vindi-

* Funding from TUF, Tekes, Digile D2I research program are gratefully acknowledged.
[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [shuhua.liu@arcada.fi]

cate the use of violence, raise cash, and legitimize their actions while demoralizing and delegitimizing others." (Cohen-Almagor 2011).

As most of the online communications is un-moderated, web forums, newsgroups and blogs become common source of different type of flames, virulent messages and rants. Hate groups use websites for sharing ideology, propaganda, link to similar sites, recruit new converts, advocate violence and threat others. Some also developed their own versions of social networking sites. Some sites which host user-created content make a virtue out of being offensive. Anytime one engages online communication, there could be a risk that he or she may be the target of ridicule or even harassment. The Simon Wiesenthal Center's Digital Hate and Terrorism project (2011), which worked with Facebook to remove hate pages, identified over 14,000 problematic websites, forums, blogs, and social media postings (Media Smarts 2016).

Exposure to online bigotry, hate and violence can have much more serious effects than simply making people feel uncomfortable or unwelcome. Studies have shown that experiencing discrimination online can cause stress, anxiety and depression (Cohen-Almagor 2011). There are also recent cases showing that hateful language can have profound impact on a community or even a major company.

Detection of hate and violence content has become an issue of increasing importance. With the constant growing and hyper-connected Internet, the amount and sophistication of harmful websites and content only proliferate. Despite efforts from web content and service providers, the spread of online hate and violence has so far outpaced efforts to remove them, and the study on automatic content filtering systems still awits to make important progress. The practical requirements for precision in identifying problematic content is very demanding. There is thus a great need for more sytemtaic efforts to address the issues in developing better performing content detection systems that can identify excessively offensive and harmful websites with higher and higher accuracy.

The availability of abundance data and the advanced developments in computing methods have brought us the fuels and new engines for building advanced content classification models (Djuric et al. 2015a). In our recent studies we have explored the potential of supervised LDA topic models, and combination of topic analysis with sentiment analysis in web content classification. In this paper we first try to understand the forms of online hate and violence. We then review related studies on hate and violence content detection, and reflect on our experience and learnt lessons with different classification methods and models for detecting hate and violence web pages (Chiang et al. 2003).

2 FORMS OF ONLINE HATE AND VIOLENCE CONTENT

Hate and violence content have proved to be much harder to detect with good accuracy than many other types of concerned content such as Adult content, Drugs and Weapon. One of the main reasons is that a clear definition of hate and violence is very difficult, not only when analyzing the content but also when investigating the potentially aggressive or offensive behavior. Different studies tend to define these notions differently and cover different forms of hate or violence. For example, some experts who track violence

in TV program define violence as the act (or threat) of injuring or killing someone, independent of the method used or the surrounding context. Some others may specifically exclude cartoon violence from their research because of its comical and unrealistic presentation. The line between hate speech and free speech is also a thin one, and different countries and culture have different levels of tolerance. It is easy to see that boundaries ambiguous and considerable content overlapping exist between Hate/Violence and other content categories (such as Religion, Cults, Occults, etc), as well as overlapping between Hate and Violence.

Although hate and violence concept and content have a wide-covering nature, they also share some common traits and purpose. Hate speech uses offensive and threatening language to expresses "discriminatory, intimidating, disapproving, antagonistic, and prejudicial attitudes" towards certain groups of people based on their gender, race, religion, ethnicity, color, nationality, disability, or sexual orientation. It is a kind of speech that demonstrates an intention to hurt, to incite harm, or to promote hatred. It is bias-motivated, hostile, malicious, offensive or illegal, and intended to injure, dehumanize, harass, intimidate, debase, degrade, or victimize the targeted groups, or to foment insensitivity and brutality against them (Cohen-Almagor 2011).

Violence and other forms of abuse towards people are most commonly understood as a behaviour pattern intended to establish and maintain power and control over a victim. Many types of inequality or imbalance of power are found to be the roots of violence and abuse. For people, violence and abuse in any form can profoundly affect health and well-being, may result in actual or potential harm to the development or dignity in the context of a relationship of responsibility, trust or power. Violence may be also targeted at larger objects such as organization, infrastructure, nature and society thus becomes big threats to our living environment and world peace.

Forms of online hate and violence are many. A good summary of forms of online Hate is presented at (Media Smarts 2016):

(1) *Websites and blogs, discussion groups maintained by hate groups* - most of which are simple screeds, but the more sophisticated ones mimic popular commercial websites in a variety of ways, with many offering audiovisual material and discussion forums and some featuring professional-looking layout and graphics. A small number of websites are designed specifically to appeal to youth and children.
(2) *Social networking:* encourage group interaction and strengthen connections between group members using sites such as Facebook and Twitter (less obvious hate content) or other special networking sites.
(3) *Multimedia content - hate music and games:* hate rock, dark metal, widespread on hate sites as well as file-sharing sites and mainstream services such as YouTube and iTunes. Hate groups also use the Web's video-sharing services to connect youth seeking guidance, support and validation to hate leaders, to appeal to youth through video games with sheer outrageousness that may provide a guilty pleasure. Games could also be used to encourage users to participate in their sites and forums. (Sureka et al. 2010)
(4) *Accidental contact:* accidental encounter to hate and violence material.
(5) *Education:* hate groups use web content as educational material to build general support for them; or, overt hate sites attacking the mainstream educational system

and call on supporters to educate their friends, families and communities about the "real" truth. Hate sites may also adopt many markers of credibility – quoting from old editions of the Encyclopedia Britannica, or selectively citing articles from reputable sources such as the Wall Street Journal. In one study one-third of hate sites denied being racist or a hate group – though often at the same time as they use more overt hate language.

(6) *Denialism, Pseudo-Science* (with flood of supposed "facts" and statistics), *Hero Narrative, Nationalism*, an appeach to *Religion, Othering* (through caricature or stereotype, name-calling, or ideology).
(7) *Scare tactics*: creating a sense of urgency around a threat is essential for hate groups to radicalize members; making use of current or controversial issues to transform fear and worry into hate.
(8) *Hate symbols:* the Nazi swastika, the KKK's burning cross, etc.
(9) *Cloaked Websites, Implicit Messages and Resonable Racism:* there are overt hate websites as well as disguised or cloaked hate sites. Overt hate sites actively promote hatred towards other groups through racist propaganda or offering hate-based communities online. Cloaked hate websites then intentionally spread hatred through more implicit and deceptive messages, appear to be legitimate sources of opinion or information, authoritative and professional at first glance, but conceal a racist agenda behind a more moderate message. Young people are vulnerable to this type of content as research shows many lack the ability to critically evaluate cloaked hate sites. Related with this is the so-called 'reasonable racism', which presents its content as political provocation or debate, relying on pseudo-science and twisted logic rather than outright expressions of hate. Studies suggest that Hate messages that are more implicit have more long-term persuasive power (Media Smarts 2016).

Violence also take many different forms. (1) Physical and psychological violence and abuse and punishment: Violence towards a person may include physical and emotional ill-treatment such as put-downs, name-calling, bullying, violent and humiliating discipline, threats and harassment, revenge and psychological terror, torture and cruel, rape, sexual abuse, neglect or negligent treatment, abandonment, inhuman and degrading treatment, maltreatment or exploitation slavery, trafficking, injury and (verbal) abuse, murder, suicide and self-mutilation, abduction, deprivation, blood, gore and other shocks. (2) There are random acts of violence and criminal activity, while there are also organised crime, gang violence, harmful traditional practices, death squads and vigilantes, extra-judicial execution, honour killings, war and terrorist activities.

Violence content are the descriptions or reports of any forms of violence. Violence language has many variations but is easy to understand and requires little context in order to present a plot. Violence in the media we consume are common – no matter in television, movies, video games, some genres of music or the Web.

3 AUTOMATIC DETECTION OF HATE AND VIOLENCE WEB CONTENT

A hate or violence site can be more formally defined as a site that uses abusive or violent language, or carries a hateful or violent message in any form of textual, visual, or audio-based rhetoric. In this article we only address the automatic detection of hate and violence content in text format, being aware that there are work on detecting hate and violence in audio and visual format as well.

3.1 Challenges in Detecting Textual Content of Hate and Violence

The most basic approaches to the identification of hate and violence content make use of predefined black-lists, a collection of keywords or sites known to be hateful or insulting. One can make a reasonably effective classifier this way and most current commercial methods make use of black-lists and regular expressions to catch bad language. However, the lists would need to be regularly updated to keep up with language changes on the web over time, while content creators trying to cleverly evade easily identifiable keywords. Some blacklist words might not be abusive in the proper context, so simple use of dictionary and keyword matching could lead to many false positives. Such an approach also falls short when encountering more subtle, less ham-fisted examples of hate and violence (Nobata et al. 2016). In addition, some insults which might be unacceptable to one group may be totally fine to another group, and thus the context of the lists becomes important. Together with the widely encompassing and sensitive nature of hate and violence, they bring many challenges for the automatic detection of hate and violence content, as summarized below.

(1) First of all, world knowledge is often needed to precisely interpret and evaluate such content, which makes hate and violence content detection not only a challenging task to automate, but also potentially very difficult for people as well (Nobata et al. 2016).
(2) Just like hate and violence may take various forms, there are also many permutations to many terms, which when coupled with the sometimes deceiving nature of presentations, makes not only automatic detection challenging but also manual checking difficult. Cloaked hate sites and ‘reasonable racism’ mean that it may often be difficult for youth to recognize hate sites when they encounter them. So detection of abusive language or general hate and violence content requires much more than simple keyword spotting (Nobata et al. 2016).
(3) Hate and violence messages may be disseminated through dedicated web sites associated with a specific group. However they may also be spread through popular sites such as Yahoo!, Twitter or Facebook where topical issues or news articles may elicit responses with offensive or abusive language (Cohen-Almagor 2011).
(4) Abusive language online can be being very noisy while in the mean time very fluent and grammatical. In other cases, Sarcasm may be present and users would post sarcastic comments in the same voice as the people that were producing abusive language (Gitari et al. 2015).

(5) Similarly, hatred against each different ethnic group may typically characterized by the use of a small set of high frequency stereotypical words. However, such words may be used in either a positive or a negative sense (Warner & Hirschberg 2012).
(6) Many application field of data analytics are benefiting from good training dataset combined with the power of new analytical methods. In hate and violence content detection however, the development of large publicly avaliable datasets is rare unitl most recent work at Yahoo. There has also been no defacto testing set with which to compare different methods (Nobata et al. 2016).

3.2 Overview of Existing Studies

One of the earliest research on hate detection concerns automatic recognition of online flames and virulent messages in emails by Spertus (1997). The Smokey system builds around a feature vector based on syntax and semantics of each sentence, and combine the vectors for the sentences within each message. It looks not only for insulting words and the context in which they are used but also for syntactic constructs that tend to be insulting or condescending, such as imperative statements. A training set of 720 email messages are used to build C4.5 decision-tree models tested on a test set of 460 messages. Her method avoids misclassifying friendly messages by looking for praise, requests, and polite speech (Spertus 1997).

After a decade, with the emergence and explosive growth of user-created content and web-based community, Yin et al. (2009) studied the detection of abusive/offensive language and harrassment on web 2.0 (discussion forums and chat rooms contained within a larger application). Their method applied supervised machine learning technique in conjunction with content features (n-gram) and sentiment features (manually developed regular expression patterns) as well as contextual features which take into account the abusiveness of previous sentences. Their study show that the identification of online harassment is feasible when TFIDF is supplemented with sentiment and contextual feature attributes.

Abbasi & Chen (2007) studied the spread of hateful messages in Dark Web forums, which are heavily used by extremist and terrorist groups for communication, recruiting, ideology sharing, and radicalization. They consider sentiment and affect analysis of Dark Web forums as a measure for the presence of hate, violence or radicalization on the internet, and performed affect analysis on U.S. supremacist and Middle Eastern extremist group forum postings. A special affect lexicon was constructed. Their analysis reveals that the Middle Eastern test bed forums have considerably greater violence intensity than the U.S. groups. In a related study, Abbasi et al. (2008) developed sentiment analysis method combines machine learning techniques with a rich set of textual feature representation to assess the sentiment polarities and affect intensities expressed in forum communications.

Chen et al. (2012) reported a best performing offensive language detection technique so far. They developed SVM based classifiers using a combination of lexical and parser features to detect offensive language in YouTube comments. The features include n-grams, automatically derived blacklists, manually developed regular expressions and

dependency parse features. The noisy raw text goes through spelling-correct and normalization before feature extraction. The definition of offensive language can be tuned by the use of a pre-set threshold. They achieve a performance on the task of inflammatory sentence detection of precision of 98.24% and recall of 94.34% (Chen et al. 2012).

Warner & Hirshberg (2012) worked on detecting hate speech (mainly racist speech) on the web using supervised models. It focuses less on abusive language but more specifically on anti-semitic hate - hateful/racist language directed towards a minority or disadvantaged group. They manually annotated a corpus of websites and user comments following a rigorous annotation procedure, adopted a supervised classification methods by first targeting certain words that could either be hateful or not, and then using Word Sense Disambiguation techniques to determine the polarity of the word. Their method performs at 0.63 F-score. Along this line, Kwok & Wang (2013) reported a preliminary study on automatic detection of Tweets against blacks.

Sood et al. (2012) were the first to use crowdsourcing to annotate abusive language. They used Amazon Mechanical Turk workers to label 6,500 internet comments as abusive or not abusive, and retained data in which a majority of the turkers agreed on the label. 9% of the comments were deemed as carrying profane words. Their study limited the task to just profanity but not other types of hate speech and abusive language. They reported an improvement in profanity detection by making use of lists as well as an edit distance metric.

Gitari et al. (2015) worked on detecting hate speech content in social media, including online forums, blogs and comments section in news reviews, using a lexicon-based approach for sentiment assessment. Their corpus are from two sources with very different orientation in terms of the target audience and presentation: (i) major data crawled on diverse dates a total of 100 blog postings (documents) from 10 different websites, 10 for each site. Most of the discussions are intellectual discourses on subtle areas such as ideology and science; (ii) one paragraph snippets of quotes relating to the Israel-Palestinian conflict. The language used here is more direct and easily discernible even by a casual reader. Their method is based on sentence-level sentiment analysis, which uses rule-based subjectivity detection to separate objective sentences from subjective sentences. They build a hate speech lexicon by extracting from the subjective sentences semantic word features that lender a sentence subjective. This domain dependence and context specific lexicon was augmented by using bootstrapping and WordNet to add hate-related verbs and dependency-type generated grammatical patterns relating to the thematic areas identified (Gitari et al. 2015).

Djuric et al. (2015b) studied hate speech detection with comment embeddings, using distributional content representation method for joint modeling of comments and words, then to train a binary classifier to distinguish between hateful or clean comments. Their approach result in hate speech detectors that outperformed bag-of-words implementation (0.8007 to 0.7889 AUC).

Most recent and substantial work on hate detection is from Yahoo (Nobata et al. 2016). They studied user comments in ”comments sections” on proprietary news and finance web articles, and developed supervised classification models for detecting abusive language in user comments based on extended sets of features including distributional se-

mantics. Their study divide hate speech detection into three key target groups: race, nationality and religion, with abusive language encompasses hate speech, profanity and derogatory language. They also developed a large corpus of user comments collected from different domains on Yahoo Finance website and annotated for abusive language, made it publicly available (Nobata et al. 2016). The data set comprises 56,280 comments containing hate speech and 895,456 clean comments generated by 209,776 anonymized users, collected and editorially labeled over a 6-month period. It is the largest hate speech data set so far reported in the literature.

Automatic detection of violence on the other hand more concentrated on violence scene detection in movies (Deniz et al. 2014) and videos, and automatic detection and forecastig of violent extremist cyber-recuritment (Scanlon & Gerber 2014).

4 REFLECTIONS ON CLASSIFICATION MODELS FOR DETECTING HATE AND VIOLENCE

Our study has been based on a training set of over 165,000 web pages and testset of a few thousands pages collected in undocumented ways, over different time period. Most probably the annotation did not follow a rigorously procedure. There were certain general guidelines about making judgement on if a web page is hate or violence or not, but the concept of hate and violence is mostly practical and would be best defined by the samples in the datasets.

Examples of Hate sites are those that denigrate a person or a group on the basis of characteristics such as race, religion or sexual orientation, sites that promote damages to a person or group, either physically or mentally, Anti Antifa, Anti Black, Anti Catholic, Anti Christian, Anti Gay, Anti Immigration, Anti Islam, Anti Semitism, Anti White, Holocaust Revisionism, Japanese Supremacy, Jihadists, Misogyny, White Supremacy. Examples of Violence sites are those Sites that contain materials such as speech, writing or images that may incite violence, Gruesome and violent images or videos, Sites that contain information on rape, harassment, snuff, bomb, assault, murder and suicide, blood and Gore, Crimes, Fights, Gore Porn, Heavy Metal (with violent and gory images/lyrics), News (with violence).

4.1 Combining Topic Similarity Analysis with Sentiment Analysis

This is a baseline approach for sentiment-aware web content detection. We choose to build binary classifiers for Hate and Violence, using balanced data sets for both Hate and Violence. We extracted topic representation for each web page using tf-idf term weighting method; extracted topic representation of each category based on Centroid summarization (Radev et al. 2004) of all web pages in each category (Hate vs non-Hate, Violence vs non-Violence); computed topic similarity between each web page and the category Hate and Violence; extracted sentiment indicators of web pages using the SentStrength (Thelwall et al. 2010; Thelwall et al. 2012) sentiment analysis techniques. We then build classifiers based on combined topic similarity and sentiment features.

SentiStrength takes a lexical approach to sentiment analysis, making use of a combination of sentiment lexical resources, semantic rules, heuristic rules and additional rules. It contains an EmotionLookupTable of 2,310 sentiment words and word stems taken from the Linguistic Inquiry and Word Count (LIWC) program (Pennebaker et al. 2003), the General Inquirer list of sentiment terms (Stone et al. 1966) and ad-hoc additions made during testing of the system. The SentiStrength algorithm has been tested on several social web data sets such as MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums. It was found to be robust enough to be applied to a wide variety of social web contexts.

We found that topic similarity based classifiers solely do not perform well, but when topic similarity and sentiment features are combined, the classification model performance is significantly improved for many web categories including Hate and Violence. The best performing validation models are RacistWhiteSupermacy reached precision 98.26%, recall 96.30%, JewRel precision 64.43%, recall 96.28%, Violence precision 93.69%, recall 82.75%. These first results was encouraging and suggested that incorporating the sentiment dimension can indeed bring much added value to web content classification. However, tets set performance on unbalanced dataset is rather disappointing, which takes us to more practical issues in our modelling work.

4.2 Practical Issues: Imbalanced Learning and Covariate Shift

Table 1 shows the distribution of Hate and Violence classes in our training and test sets. Highly skewed class distributions like ours are not uncommon in real world applications. However, learning algorithms usually assume that the ratios of each class are close to equal and the errors associated with each class have the same cost. With imbalanced dataset, the cost gets skewed in favor of the majority class, and models built with imbalanced dataset will cause the underrepresented class to be overlooked or even ignored. Thus, a common belief is that we should balance the class prevalence before training a classifier to improve performance.

Table 1. Datasets for Hate and Violence

Classes	**Training set**	**Test set**
Hate	1733 (2.58%)	184 (5.84%)
Violence	1400 (1.92%)	135 (4.37%)
Complete data set	67212/73107 (Hate/Violence)	3153/3086 (Hate/Violence)

Solutions for imbalanced learning thus include sampling based, cost sensitive methods and active learning methods. Resampling tries to achieve dataset balance artificially so that the prevalence of the minority class is enriched. In general two sampling strategies

can be employed for balancing the classes: oversampling (adding instances to the minority class) and under-sampling (removing instances from the majority class). Cost based learning is based on cost sensitive modeling that weights the costs of misclassifying the majority class (false negatives) and the minority class (false positives) separately. By training the learner to minimize overall cost it gives more incentive for more true positives. There are also cost-sensitive ensembles in addition to cost-sensitive functions incorporated directly into classification methods (He & Garcia 2009).

A different view of the class imbalance problem is that poor performance is caused by there not being enough patterns belonging to the minority class, not by the ratio of positive and negative patterns itself. Generally when there is enough data, the "class imbalance problem" doesn't arise (He & Garcia 2009). Thus the essence of issues with imbalanced learning is not only the disproportion of positive and negative classes, but also the poor representativeness of the minority class due to its small size.

The same issue could happen to majority class as well. For example, in many classification tasks on web scale, positive and unlabeled data are widely available, whereas collecting a reasonable and representative sampling of the negative examples could be challenging, because the negative data set, as the complement of the positive one, should uniformly represent the universal set excluding the positive class, but such probability distribution can hardly be approximated (Yu et al. 2004).

To handle class imbalance issue we adopted the under-sampling strategy and experimented with three sampling strategies: natural distribution (native prevalence), fully balanced (about 50/50) and less balanced (20/80) distributions. Our experiments indicate that artificial balancing of the classes brings a positive effect on the model performance. Although all models performance degrades as target prevalence decreases, enriching the target class prevalence during training helped improve performance on both the validation and test result (with the target at its native prevalence), especially evident with validation results. The changing of model types does not make much difference. We also tested cost-sensitive models coupled with threshold control, which again help to improve precision on positive class. However, overall the effect of threshold control is still rather limited (around max 3% gain in precision, with tradeoff on recall), can't bring performance up in a significant way.

When a dataset is artificially balanced, it often implies that there is close to equal prior probability of positive and negative patterns. When that is not the case, the model could also make poor predictions by over-predicting the minority class. In practice it often happens that the training set and test set not only both have highly skewed class priors, but the class distribution also very different from each other. This is referred to as Covariate Shift issue, which often cause model performance degradation in the test set, which is especially obvious with Naive Bayes and Logistic Regression based classifiers (Bickel et al. 2007). To address this issue, we tried to incorporate word semantics making use of relevant concepts of Hate and Violence in ConceptNet. We considered the effect of ConceptNet terms on topic words, but we wasn't able to find positive effect on the classifier performance. We continue to search for better approach to handle the covariate shift issue with our application.

4.3 Supervised LDA Topic Modelling

Although addressing imbalanced learning helped us to enhance the test set performance with a big margin, there is still much room for improvements. We turned to LDA topic modeling which offres a more sophisticated treatment of topic extraction and text classification. LDA topic models are probabilistic models for discovering the hidden thematic structure in large document collections based on a hierarchical Bayesian analysis method (Blei et al. 2003). With supervised LDA method, extracted topic features instead of word features can be used for learning with much reduced dimensionality (Blei & McAuliffe 2007).

Topics are defined as a distribution over a fixed vocabulary of terms; documents are defined as a distribution over topics; with the distributions all automatically inferred from analysis of the text collection. By discovering patterns of word use and connecting documents that embrace similar patterns, topic models prove to be a powerful technique for finding topic structure in text collections. Topic models can help us answer questions such as what topics are contained in the document collection, what subject matters each document discusses, and how similar one document is to another.

Our results showed that supervised LDA topic modeling based classifiers outperform all our earlier models on test dataset, for both Hate and Violence. However, recall is the key improvement (92%); very close on precision level (52%), which possibly gives room to bring up the precision level at trade-off of recall level. On combined detection of Hate and Violence, our best model with balanced traning set reached new performance levels 66-72% precision, 85-93% recall.

5 SUMMARY

Studies on automatic detection of hate and violence web content have focused on a few different forms of hate and violence. Some studies focus on detecting profanity, some focus on detecting abusive language, some focus on hate speech directed to a particular ethnic group. Different works may tackle specific aspects of abusive language, define the term differently, or apply it to specific online domains only (Twitter, online forums). Sometimes different studies spread across several overlapping fields, however overall they are fragmented. Our study on hate and violence detection has a more general and broader topic and online domain coverage.

Many analytics applications are benefiting from good training dataset combined with the power of new analytical methods. In hate and violence detection, nearly all previous work uses different evaluation sets. New efforts just emerged to provide publicly accessible dataset for hate detection. Our data set is in decent size, but could certainly be enriched by the pulicly available new dataset on hate content. Our research could also be better benchmarked by testing on a common evaluation set.

Rethinking our strategies for building classification models for detecting hate and violence web pages, some natural extensions include integrating LDA topic modeling with similarity-based approach, modeling using positive and unlabeled data, classifier en-

semble (random forest, gradient boosting), as well as integration of text based and image based classifiers. Fast developments and many breakthroughs in machine learning methods and natural language technology are enabling new methods for content representation and understanding that could help us develop better content detection systems for the detection of violence, intolerance and hateful web content. It will also make sense to look at the raw data input to text analysis system, to devote more efforts in data collection using Amazon Mechanic Turk, or to treat content detection and classification as a paragraph or sentence classification tasks instead of only a document classification task.

REFERENCES

Abbasi, A. & Chen, H. 2007, Affect intensity analysis of Dark Web forums, in: *Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics*, New Brunswick, NJ, pp. 282-288.

Abbasi, A., Chen, H., & Salem, A. 2008, Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums, in: *ACM Transactions on Information Systems (TOIS)*, Vol. 26, Iss. 3.

Bickel, S., Brückner, M., & Scheffer, T. 2007, Discriminative learning for differing training and test distributions, in: *Proceedings of the 24th International Conference on Machine learning*, New York: ACM Press, pp. 81-88.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003, Latent dirichlet allocation, in: *Journal of Machine Learning Research*, No. 3, pp. 993-1022.

Blei, D. M. & McAuliffe, J. D. 2007, Supervised Topic Models, in: *Neural Information Processing Systems 21*.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. 2012, Detecting offensive language in social media to protect adolescent online safety, in*: 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Confernece on Social Computing (SocialCom)*, IEEE Press, pp. 71–80.

Sureka, A., Kumaraguru, P., Goyal, A., & Chhabra, C. 2010, Mining YouTube to Discover Extremist Videos, Users and Hidden Communities, in: Cheng, P.-J., Kan, M.-Y., Lam, W., Nakov, P. (Eds.) *Information Retrieval Technology*, LNCS, Vol. 6458, Springer, pp. 13-24.

Chiang, C., Gerstenfeld, P., & Grant, D. 2003, Hate Online: A Content Analysis of Extremist Internet Sites, in: *Analyses of Social Issues and Public Policy*, Vol. 3, No. 1, pp. 29-44.

Cohen-Almagor, R. 2011, Fighting Hate and Bigotry on the Internet, in: Policy and Internet, Vol. 3, No. 3.

Deniz, O., I. Gracia, I., S., Bueno, G., & Kim, T.-T. 2014, Fast Violence Detection in Video Games, in: *Proceedings of the 9th International Conference on Computer Vision Theory and Applications*.

Digital Hate and Terrorism Project. 2011, Simon Wiesenthal Center.

Djuric, N., Wu, H., Radosavljevic, V., Grbovic, M., & Bhamidipati, N. 2015a, Hierarchical neural language models for joint representation of streaming documents and their content, in: *Proceedings of the 24th International Conference on World Wide Web*, ACM Press, pp. 248-255.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. 2015b, Hate speech detection with comment embeddings, in: *Proceedings of the 24th International Conference on World Wide Web*, ACM Press, pp. 29-30.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. 2015, A Lexicon-based Approach for Hate Speech Detection, in: *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 10, No. 4, pp.215-230

He H. & E. A. Garcia, E. A. 2009, Learning from Imbalanced Data, in: *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, pp. 1263-1284.

Kwok, I. & Wang, Y. 2013, Locate the Hate: Detecting Tweets against Blacks, in: *Proceedings of the 27th National Conference on Artificial Intelligence (AAAI)*, pp. 1621-1622.

Levmore, S. & Nussbaum, M. (Eds.). *The Offensive Internet: Speech, Privacy, and Reputation*, Cambridge, Massachusett: Harvard University Press.

Media Smarts. 2016. Deconstructing Online Hate. Accessed 6.6.2016. Published 2016. http://mediasmarts.ca/online-hate/deconstructing-online-hate

Nobata. C., Tetreault, J., Thomas, A. Mehdad, Y., & Chang, Y. 2016, Abusive Language Detection in Online User Content, in: *Proceedings of the 25th International Conference on World Wide Web*, pp. 145-153.

Pennebaker, J. W., Mehl, M. R., & Niederhofer, K. G. 2003, Psychological Aspects of Natural Language Use: Our Words, Our Selves, in: *Annual Review of Psychology*, Vol. 54, pp. 547-577.

Radev, D., Jing, H., Styś, M., & Tam, D. 2004, Centroid-based summarization of multiple documents, in: *Information Processing & Management*, Vol. 40, Iss. 6, pp. 919– 938.

Scanlon, J. R. & Gerber, M. S. 2014, Automatic detection of cyber-recruitment by violent extremists, in: *Security Informatics*, Vol. 3, Iss. 1.

Sood, S. O., J. Antin, J., & E. F. Churchill, E. F. 2012, Using Crowdsourcing to Improve Profanity Detection, in: *AAAI Technical Report SS-12-06 Wisdom of the Crowd*, pp. 69-74.

Spertus, E. 1997, Smokey: Automatic Recognition of Hostile Messages, in*: Proceedings of the 8th Annual Conference on Innovation Application of AI (IAAI)*, pp. 1058–1065.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. 1966, *General Inquirer: A Computer Approach to Content Analysis*.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. 2010, Sentiment strength detection in short informal text, in: *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 12, pp. 2544– 2558.

Thelwall, M., Buckley, K., & Paltoglou, G. 2012, Sentiment strength detection for the social Web, in: *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 1, pp. 163-173.

Warner, W. & Hirschberg, J. 2012, Detecting Hate Speech on the World Wide Web, in: *Proceedings of the Second Workshop on Language in Social Media*, pp. 19-26.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. 2009, Detection of Harassment on Web 2.0. in: *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*.

Yu, H., Han, J., & Chang, K. C.-C. 2004, PEBL: Web Page Classification without Negative Examples, in: *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, Iss. 1, pp.70-81.

Evaluating Confidence Intervals for ELM Predictions

Anton Akusok[i,ii], Yoan Miche[iii,iv], Kaj-Mikael Björk[v], Rui Nian[vi], Paula Lauren[vii], Amaury Lendasse[viii]

Abstract

This paper proposes a way of providing more useful and interpretable results for ELM models by adding confidence intervals to predictions. Unlike a usual statistical approach with Mean Squared Error (MSE) that evaluates an average performance of an ELM model over the whole dataset, the proposed method computed particular confidence intervals for each data sample. A confidence for each particular sample makes ELM predictions more intuitive to interpret, and an ELM model more applicable in practice under task-specific requirements. The method shows good results on both toy and a real skin segmentation datasets. On a toy dataset, the predicted confidence intervals accurately represent a variable magnitude noise. On a real dataset, classification with a confidence interval improves the precision at the cost of recall.

Keywords: extreme learning machines, confidence, confidence interval, regression, image segmentation, skin segmentation, classification, interpretability, big data

1 INTRODUCTION

Extreme Learning Machines (ELM) (Huang et al. 2012; Huang et al. 2004; Huang 2015) are fast (van Heeswijk et. al 2011) and robust (Miche et al. 2010; Miche et al. 2011) methods of training feed-forward networks, which have the universal approximation property (Huang et al. 2006) and have numerous applications in regression (Yu et

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [anton.akusok@arcada.fi]
[ii] The University of Iowa, USA, College of Engineering
[iii] Nokia Solutions and Networks Group, Finland, [yoan.miche@nokia-bell-labs.com]
[iv] Aalto University, Finland, School of Science
[v] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [bjorkpau@arcada.fi]
[vi] Ocean University of China, School of Information Science and Engineering, [nianrui_80@163.com]
[vii] Oakland University, USA, [paula_lauren@hotmail.com]
[viii] The University of Iowa, USA, College of Engineering, [amaury-lendasse@uiowa.edu]

al. 2013; Miche et al. 2008; Cambria et al. 2013) and classification (Akusok et al. 2015) problems. They are an active research topic with numerous extensions and improvements proposed over the last decade.

ELMs are powerful non-linear methods, but they share one common drawback of non-linear methods in practical applications, which is a non transparency of results (predictions). A prediction made by a linear model from input data is easily explained and interpreted by observing the coefficients at input data features. Results which have an explanation are easier to trust and apply for people outside a Machine Learning field. Non-linear models lack such transparency, so their results are hard to be trusted, and thus non-linear methods (including ELM) are sometimes denied despite a supreme performance compared to linear methods.

This paper proposes a way of providing more useful and interpretable results for ELM models by adding confidence intervals (Shang & He 2015; Lendasse et al. 2005; Pouzols et al. 2010; Guillen et al. 2008) to predictions. Unlike a usual statistical approach with Mean Squared Error (MSE) (Bishop 2006) that evaluates an average performance of an ELM model over the whole dataset, the proposed method computed particular confidence intervals for each data sample. These intervals are small for samples on which a model is accurate, and large for samples where a model is unstable and inaccurate. A confidence for each particular sample makes ELM predictions more intuitive to interpret, and an ELM model more applicable in practice under task-specific requirements to precision and recall of predictions.

The next section 2 introduces the method of input-specific confidence intervals. The experimental section 3 presents the examples of confidence intervals on artificially made toy dataset and on a real image segmentation task. In the conclusion, section 4 the method is summarised, and further research directions are discussed.

2 METHODOLOGY

Confidence intervals are estimated boundaries of a stochastic output sample for a given input sample and confidence level, in a regression or classification task. They provide a measure of confidence for a prediction result of an ELM. This information is practically important in many ELM applications, and is useful in complex systems which utilize ELM as their part.

A simple way of estimating confidence interval of ELM predictions is to use Mean Squared Error (MSE), which is a variance of error between model predictions and true output values. But this method provides constant confidence intervals for the whole dataset, while predictive performance of ELM may vary depending on the input. More useful confidence intervals are defined in the input space, as described hereafter.

Confidence intervals are estimated from the variance (or standard deviation), however predictions of a single ELM are deterministic. To obtain stochastic predictions, this work considers a family of ELM models. Input weights of these ELMs are randomly sampled, but each model has the same parameters including random weights distribu-

tion, projection function and network structure. As ELM is a very fast training method, hundreds to millions of ELMs can be trained in a few minutes (depending on training data and model size), providing adequately precise estimation of outputs distribution of an ELM model family.

2.1 Confidence Interval for Regression

A data set is a limited set of N samples $\{x_i, t_i\},\ i \in [1, N]$ which represents an unknown projection function $F : X \rightarrow T$. An ELM approximates that function F by a smooth function f such as $f(x_i) = y_i = t_i + \epsilon$. The noise ϵ comes from an imperfect approximation of the true projection function, noise in the dataset, and uncertainty of the dataset itself.

An assumption is made that a model prediction y_i is normally distributed $\mathcal{N}(\mu_i, \sigma_i^2)$ where $\sigma_i = \sigma(x_i)$ is defined in the input space. The the confidence intervals for an input sample x_i are computed from $\sigma(x_i)$ at the desired confidence level. However, evaluating $\sigma(x)$ for arbitrary x is complicated because the dataset input samples do not cover all input space.

In fact, the $\sigma(x)$ needs to be evaluated only for the given input points, not the whole input space. These evaluations are obtained using ELM models, which cover the whole input space (an ELM produces an output for any input sample) and can evaluate $\sigma(x)$ for any given input sample directly.

The standard deviation $\sigma(x_i)$ is evaluated by training multiple ELMs with the same network structure but different randomly sampled hidden layer weights. The obtained $\sigma(x_i)$ is influenced by a local data outputs distribution and a model structure $\sigma_1(x_i) = \sigma^{data}(x_i) + \sigma^{model}(x_i)$. Unfortunately, the model structure influence is dominant and cannot be removed by training a large number of ELM models, see Figure 1.

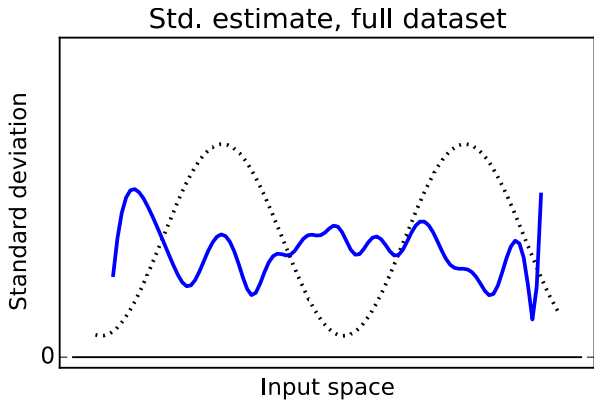


Figure 1. Estimated standard deviation of data (solid line) with a family of 10,000 ELM models, and true standard deviation of data (dotted line).

The following method is proposed to remove model influence $\sigma^{model}(x_i)$. Each model in the ELM family is trained again, but with a smaller random subset of data samples $\{x_j, t_j\},\ j \in [1, M < N]$. The model component of $\sigma(x)$ will be the same because the

ELM models are the same, and the data component will increase because with less training samples ELMs have worse fit and larger variance of predictions. The result will be $\sigma_2(x_i) = (1+\beta)\sigma^{data}(x_i) + \sigma^{model}(x_i)$ with some positive $\beta > 0$ (see Figure 2). The model-independent estimation is obtained as $\sigma(x_i) = \sigma_2(x_i) - \sigma_1(x_i) = \beta\sigma^{data}(x_i) \sim \sigma^{data}(x_i)$ (see Figure 3).

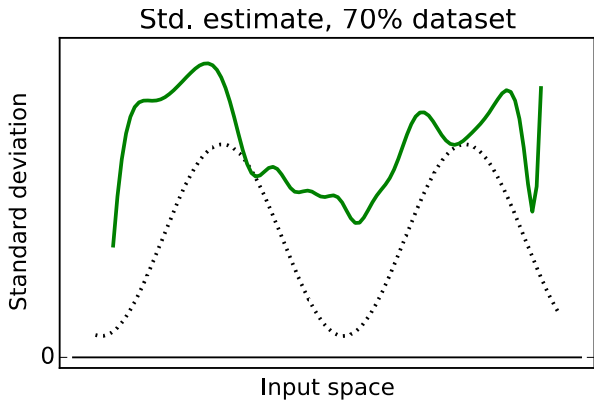


Figure 2. Estimated standard deviation of data (solid line) with a family of 10,000 ELM models, using 70% random training samples.

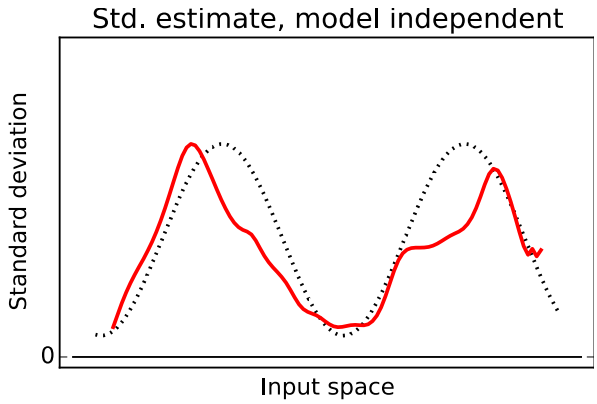


Figure 3. Estimated standard deviation of data (solid line) with a family of 10,000 ELM models, as a difference between ELM model family trained on a full dataset and the same family trained on randomly selected 70% training data (each).

The scale α of an estimate $\sigma(x_i) \sim \sigma^{data}(x_i) = \alpha\sigma(x_i) = \alpha\big(\sigma_2(x_i) - \sigma_1(x_i)\big)$ is not defined. It is obtained using a validation dataset, and a desired confidence level. For confidence level c, α is adjusted such that $1-c$ validation samples are outside the interval $y_i \pm \alpha(\sigma_2(x_i) - \sigma_1(x_i))$. The computed α is then used for calculating the confidence intervals for test data samples.

3 EXPERIMENTAL RESULTS

3.1 Artificial Dataset

An artificial dataset (Figures 4) has one-dimensional input and target data, for the ease of visualization. The data is a sum of two sine functions. Noise has been added to data samples, with varying magnitude. Because the noise is added artificially, the exact 1σ and 2σ confidence boundaries are known, and can be compared with the estimated boundaries by the proposed method. The method uses 1000 training, 9000 validation and 1000 test samples.

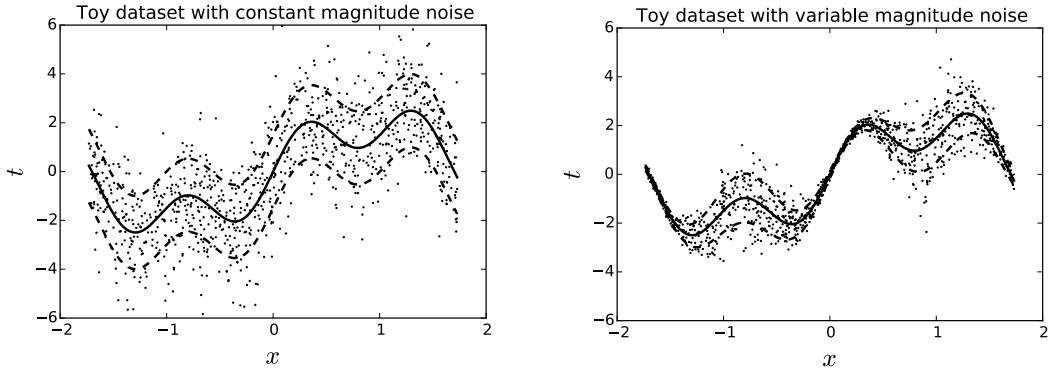


Figure 4. Artificial datasets with added constant magnitude noise (right) and variable magnitude noise (left). Dots are training samples, solid line is the true function, and dash lines show 1σ *confidence boundaries.*

The confidence boundaries are shown on Figure 5 (top) for variable magnitude noise and Figure 5 (bottom) for constant magnitude noise. All experiments train 1000 different OP-ELM models with 25 hidden neurons, which takes between 37 and 40 seconds on 1.4GHz dual-core laptop using a toolbox from (Akusok et al. 2015).

As a performance measure, an integral of absolute difference between two boundaries is evaluated, using the test samples. This integral is divided by an integral of the true boundary.

Delta is given by $$\delta = \frac{\sum_{x_{\text{test}}} |\sigma(x) - \hat{\sigma}(x)|}{\sum_{x_{\text{test}}} \sigma(x)} \times 100\%$$

With variable noise, delta of an ELM-estimated boundary is $\delta^{ELM} = 9\%\ldots11\%$, while the delta of MSE boundary is $\delta^{MSE} = 50\%$. For constant noise, $\delta^{ELM} = 12\% \ldots 16\%$ and an MSE provides almost perfect boundary estimation with $\delta^{MSE} < 2\%$. ELM does not provide a smooth boundary in case of a constant noise, because while noise is being constant, the data itself is highly varying which is reflected by an ELM estimation. Also, ELM-estimation is not accurate at the edges of a dataset.

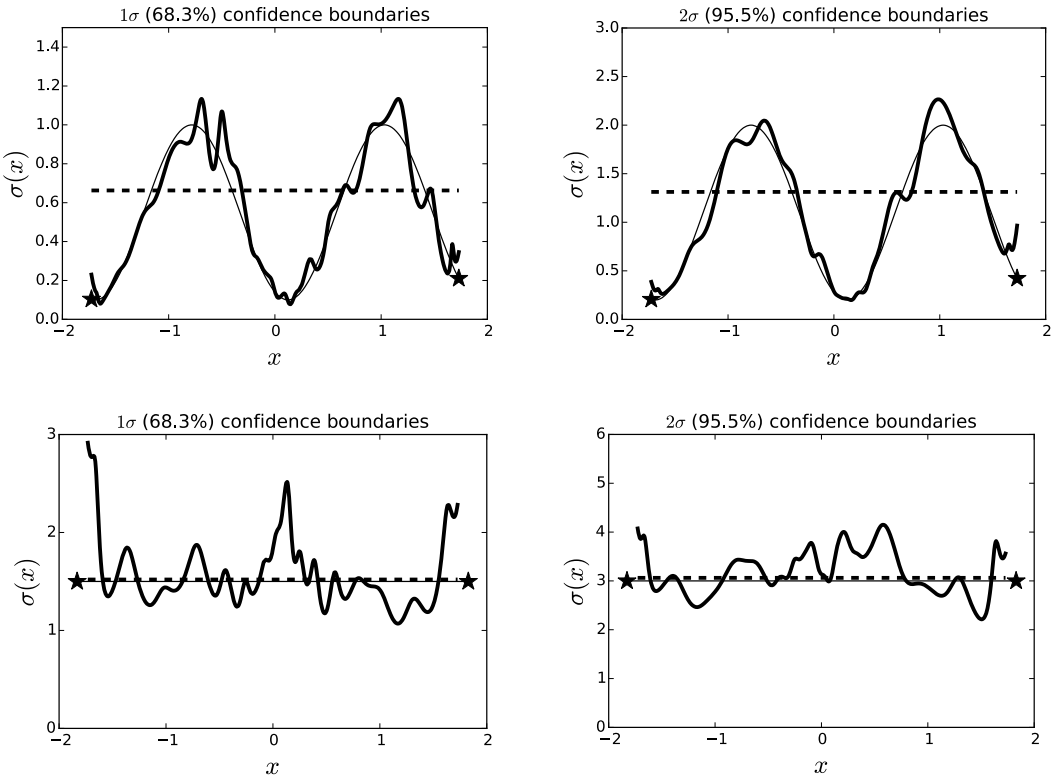


Figure 5. Boundaries for 1σ and 2σ confidence levels on toy dataset with variable (top) and constant (bottom) magnitude noise. True shape of noise magnitude is shown by a thin curve ending with stars, ELM-estimated boundary by a thick curve, and MSE boundary by a dash line.

3.2 Skin Color Dataset

Confidence intervals for ELM predictions are tested on a Face/Skin Detection dataset (Phung et al. 2005), a useful benchmark (Swaney et al. 2015) Big Data dataset. It includes 4000 photos of people under various real-world conditions, as well as manually created masks for faces and skin. The dataset is split into 2000 training and 2000 test images.

Confidence intervals are estimated for a simple task of classifying a pixel into skin/non-skin, based on its RGB color. 500 random skin and non-skin pixels are taken from each training image; 500 training set images are used for training and 1500 for validation. All pixels of a single test set image are added as test samples, for which the classification and confidence intervals are computed. A total of 1000 ELM models are built for confidence interval estimation.

The results are shown on Figure 6. An original image is split into skin and non-skin with a good accuracy using a simple threshold, however some parts are misclassified. Thresholding pixels with more than 68% confidence (roughly 1σ) significantly reduces the recall, but provides almost perfect precision. Increasing a threshold to 95% confidence (or 2σ) provides perfect precision at a cost of even smaller recall.

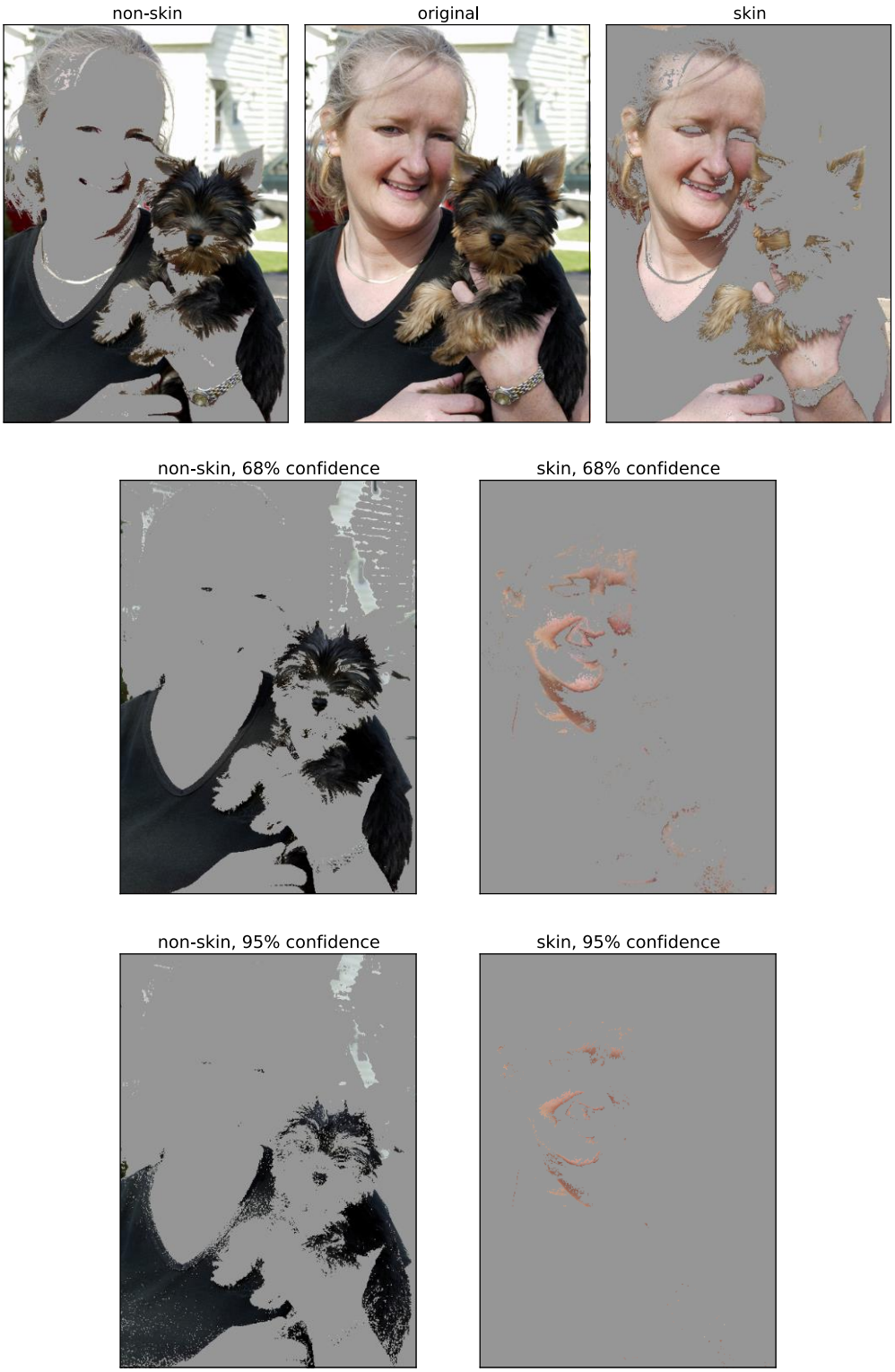


Figure 6. Skin segmentation by pixel classification with ELM based on their RGB color values, using a threshold. Top row: original image and thresholded skin/non-skin pixels. Middle row: thresholded pixels with predicted values larger than their 68% confidence intervals. Bottom row: thresholded pixels with predicted values larger than their 95% confidence intervals. Thresholding with confidence interval improves the precision at a cost of recall.

Another experiment is performed in a similar setup, but a color wheel image is used instead of a test picture. The confidence intervals for different colors are shown on Figure 7. Confidence intervals are organized according to color values --- they are small for skin colors and clearly non-skin colors, and are large in between these two. An MSE confidence interval would be constant over all colors in a color wheel, providing less accurate results.

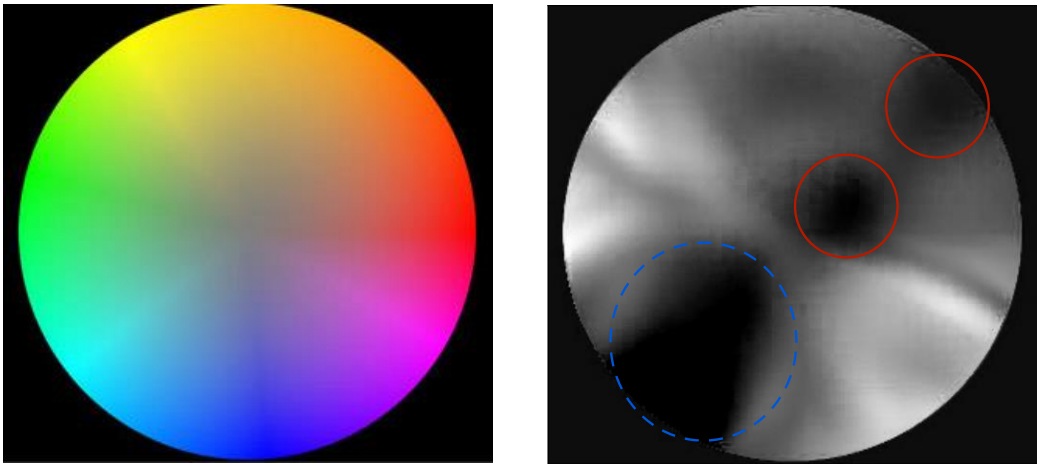


Figure 7. Skin color confidence interval map, shown on a color wheel. Small confidence intervals (black) show high confidence regions --- for skin (drawn by red solid circles) and non-skin (drawn by a blue dashed circle). Large confidence intervals (white) show low confidence for skin/non-skin classification of that color.

4 CONCLUSIONS

A method for evaluating input-dependent confidence intervals for ELM model is proposed in the paper. It is based on estimation of a standard deviation of output (target) noise for a fixed ELM model trained on a different subsets of the training set. Then the confidence intervals are scaled using a validation set, and evaluated for the given test input samples. The method can compute confidence intervals even for a high dimensional inputs, because they are computed for the given test samples and not for the whole input space.

The method shows good results on both toy and real datasets. On a toy dataset, the predicted confidence intervals accurately represent a variable magnitude noise. On a real dataset, classification with a confidence interval improves the precision at the cost of recall.

Future work on confidence intervals for ELM will be focused on computing confidence intervals for classification tasks, in a classification-specific way without value threshold. Other directions are creating more smooth confidence intervals for different models, and formulating rules for choosing parameters of the proposed method that ensure good and stable results in any application scenarios. Targeting Big Data is also an important future works topic, as the confidence interval estimations will add more value to ELM ability of handle Big Data.

REFERENCES

Akusok, A., Björk, K.-M., Miche, Y., & Lendasse, A. 2015, High-Performance Extreme Learning Machines: A Complete Toolbox for Big Data Applications, in: *IEEE Access*, Vol. 3, pp. 1011-1025.

Akusok, A., Miche, Y., Karhunen, J., Björk, K.-M., Nian, R., & Lendasse, A. 2015, Arbitrary Category Classification of Websites Based on Image Content, in: *IEEE Computational Intelligence Magazine*, Vol. 10, No. 2, pp. 30-41.

Bishop, C. M. 2006, *Pattern Recognition and Machine Learning*, Springer.

Cambria, E. et al., 2013, Extreme Learning Machines, in: *IEEE Intelligent Systems*, Vol. 28, No. 6, pp. 30-59.

Guillen, A., Sovilj, D., Lendasse, A., Mateo, F., & Rojas, I. 2008, Minimising the delta test for variable selection in regression problems, in: *International Journal of High Performance Systems Architecture*, Vol. 1, No. 4, pp. 269-281.

Huang, G.-B. 2015, What are Extreme Learning Machines? Filling the Gap between Frank Rosenblatt's Dream and John von Neumann's Puzzle, in: *Cognitive Computation*, Vol. 7, No. 3, pp. 263-278.

Huang, G.-B., Zhou, H., Ding X., & Zhang, R. 2012, Extreme Learning Machine for Regression and Multiclass Classification, in: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, Vol. 42, No. 2, pp. 513-529.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. 2004, Extreme Learning Machine: a New Learning Scheme of Feedforward Neural Networks, in: *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks*.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. 2006, Extreme learning machine: theory and applications, in: *Neurocomputing*, Vol. 70, No. 1, pp. 489-501.

Lendasse, A., Ji, Y., Reyhani, N., & Verleysen, M. 2005, LS-SVM Hyperparameter Selection with a Nonparametric Noise Estimator, in: *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005*, LCNS, Vol. 3697, Springer, pp. 625-630.

Miche, Y., Bas, P., Jutten, C., Simula, O., & Lendasse, A. 2008, A methodology for building regression models using extreme learning machine: OP-ELM, in: *Proceedings of European Symposium on Artificial Neural Networks (ESANN)*, Brudges, pp. 23-25.

Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten C., & A. Lendasse, A. 2010, OP-ELM: optimally pruned extreme learning machine, in: *IEEE Transactions on Neural Networks*, Vol. 21, No. 1, pp. 158-162.

Miche, Y., van Heeswijk, M., Bas, P., Simula, O., & and Lendasse, A. 2011.TROP-ELM: A double-regularized ELM using LARS and Tikhonov regularization, in: *Neurocomputing*, Vol. 74, No. 16, pp. 2413-2421.

Phung, S. L., Bouzerdoum, A., & Chai, D. 2005, Skin segmentation using color pixel classification: analysis and comparison, in: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 1, pp. 148-154.

Pouzols, F. M., Lendasse A., & Barros, A. B. 2010, Autoregressive time series prediction by means of fuzzy inference systems using nonparametric residual variance estimation, in: *Fuzzy Sets and Systems*, Vol. 161, No. 4, pp. 471-497.

Shang, Z. & He, J. 2015, Confidence-weighted extreme learning machine for regression problems, in: *Neurocomputing*, Vol. 148, pp. 544-550.

Swaney, C. Akusok, A. Björk, K.-M., Miche, Y., & Lendasse, A. 2015, Efficient skin segmentation via neural networks: HP-ELM and BD-SOM, in: *Procedia Computer Science*, Vol. 53, No. 1, pp. 400-409.

van Heeswijk, M, Miche, Y., Oja E., & and Lendasse, A. 2011, GPU-accelerated and parallelized ELM ensembles for large-scale regression, in: *Neurocomputing*, Vol. 74, No. 16, pp. 2430-2437. Q.

Yu, Q., Miche, Y., Eirola, E., van Heeswijk, M., Severin, E., & Lendasse, A. 2013, Regularized extreme learning machine for regression with missing data, in: *Neurocomputing*, Vol. 102, pp. 45-51.

Predicting Systemic Financial Stress

Markus Holopainen[i], Peter Sarlin[ii,iii]

Abstract

The global financial crisis has sparked interest in new techniques and methods for assessing systemic risk. While early-warning literature on binary classification of pre-crisis periods has been plentiful, regression-based forecasting of a continuous systemic stress indicator remains largely unexplored. This short paper outlines the opportunities and challenges for early-warning regression and puts forward a suitable set-up as a modeling framework.

Keywords: systemic risk, early-warning, regression

1 INTRODUCTION

The still ongoing financial crisis and global economic turmoil has stressed the need for theoretical and empirical work in deriving new approaches and techniques for measuring systemic risk. This particularly concerns early-warning models, which commonly use classification techniques to assess whether an observation belongs to a (pre-)crisis state or not. However, literature regarding its regression counterpart has been sparse, despite that it basically attempts to tackle the same prediction problem, yet at a more detailed level. In this short paper, we outline the opportunities in and the methodology for using regression for continuous early-warning models. Section 2 presents the general methodology for the regression case, while Section 3 presents the data and methods to be used. Section 4 concludes.

[i] Arcada University of Applied Sciences, Finland, Risklab, [markus@risklab.fi]
[ii] Arcada University of Applied Sciences, Finland, Risklab, [peter@risklab.fi]
[iii] Hanken School of Economics,Finland, Department of Economics

2 METHODOLOGY

This section firstly presents the challenges in regression of a continuous stress indicator in relation to the commonly used classification between binary pre-crisis and tranquil events. Secondly it outlines appropriate evaluation procedures for regression in this setting.

2.1 Classification vs. Regression

Within the area of systemic risk analytics, we focus on early-warning modeling as a general-purpose problem. Commonly, the aim is to classify between pre-crisis periods (1) and tranquil periods (0), where pre-crisis periods are usually defined as e.g. 8 quarters before a crisis event. The crisis events are usually defined through thresholds on a continuous stress index. Based upon this setting, there could be two separate approaches to tackle the problem: (i) classification for binary early-warning models and (ii) regression for continuous early-warning modeling. The former focuses on the use of classification techniques for deriving early-warning models, whereas the latter would focus on forecasting (with a fairly long horizon) a continuous stress index. Yet, the latter is to a large extent unexplored in the literature. Despite basically tackling the same problem, the advantage of the latter is that we need not assume crises to be identical binary events, but can instead account for the size of the modeled systemic events directly in the learning algorithm. This is a desirable property especially for panel data, where the magnitude of country-specific events thus can be taken into consideration.

This approach has obvious implications for evaluation and modeling. In evaluation of the classification case, one commonly studies the total sums of correctly vs. incorrectly classified observations, and derives evaluation measures based on these values. For regression, models require evaluation based on certain criteria, which usually measure the deviations of the regressed output compared to the actual values. These deviations are commonly referred to as noise, which descends from randomness and/or variables not accounted for, and is generally unavoidable. As with classification, the goodness-of-fit is thus subject to field-specific experience.

For modeling, the continuous target indicator along with a suitable forecast horizon needs to be carefully chosen. In addition, the methods chosen for the task need to be able to output continuous regressed values, instead of the binary and/or probabilistic output of methods in the classification case.

2.2 Modeling Objective and Evaluation

The starting point for any modeling task is to define the ultimate objective of the model and the evaluation criteria, based on which the model performance is measured. In this paper, broadly speaking, we wish to fit a number of methods to a continuous response variable based on continuous predictors by means of regression. The fit of each of these models are then evaluated based on measures indicating the deviations of the output compared to the actual values, averaged over all available observations.

In line with the early-warning task of classification of binary pre-crisis/tranquil events, a proper forecasting horizon needs to be determined. Due to the continuous nature of the output, a forecasting horizon, with the length of one time unit (based on the data available) needs to be chosen. In line with Holopainen & Sarlin (2015), we propose experimentation with three forecast horizons, namely 4, 8, and 12 quarters. These horizons reflect the viewpoint of a policymaker, to which e.g. detecting early-warning signals with a horizon shorter than one year may be too late for further in-depth investigation and the implementation of macroprudential tools.

A crucial modeling and evaluation decision relates to which portion of the available data to use for training, and which portion to use for evaluation. A sought-after property of any model is its ability to generalize, i.e. how well the model performs on previously unseen data. If the entire data set is used both for training and evaluation, one will obtain results which refer to in-sample performance and do not reveal much about the model's true predictive power. Additionally, more complex non-linear methods may find patterns in the data which are characteristic for that particular training set, and the resulting model may not predict well on another realization of the data set, given the same predictors. The phenomenon of overfitting is commonplace with non-linear methods when their complexity settings are high, but also with, for instance, a linear polynomial regression model if polynomials of unnecessary high order are added to the model, with the intention to increase performance. A properly constructed evaluation exercise takes these features into consideration and attempts to minimize overfit and improve out-of-sample performance.

For many decades the procedure of choice to evaluate models, whilst accounting for their generalization abilities, was cross-validation (see Stone (1977)), of which the so-called *k*-fold variant randomly splits the sample into *k* folds of similar size. Each fold is in turn treated as an out-of-sample subset, the model is trained using the remaining *k*-1 folds and evaluated on the out-of-sample subset. This is repeated for all folds and all out-of-sample predictions from the folds are collected at the end and evaluated as a whole. However, a critical consideration regarding early-warning modeling is that data most likely will exhibit dependencies in the cross-sectional as well as in the time dimension. Although the literature has investigated techniques to decrease the effects of dependence in cross-validation (see e.g. Arlot & Celisse (2010)), a preferable approach is to truly account for the time dimension by utilizing only historical information as training data for each prediction. With regard to the real-time analysis perspective, we thus propose the use of the recursive exercise presented in Holopainen & Sarlin (2015) for this setting. The recursive exercise derives a new model at each consecutive point in time, based on the historical data then available. The idea is to define a starting point in time for the recursion, ideally before some major event affecting the financial system had occurred, and to proceed onwards deriving new models at each point in time. This enables testing whether the models would have provided means of predicting future events, and function adequately as early-warning indicators.

As most machine learning methods involve free parameters, which affect their complexity, we also propose to use the recursive exercise in combination with a grid search for model selection.

3 DATA AND METHODS

This section discusses the data and the methods proposed for continuous forecasting of systemic financial stress.

3.1 The financial stress index

Due to the regression case not having the property of a binary-type output, the choice of which stress index to forecast is largely relevant. While it is not viable to assume that a single numerical statistic is able to fully characterize a system as complex as systemic risk, the potential in studying a properly constructed composite index should not be underestimated. Such an index may be able to assess the instabilities in the whole financial system, as well as to better describe different levels of risk, due to its property of being continuous.

For this application we replicate an existing financial stress index (FSI), specifically the Sovereign Composite Indicator of Systemic Stress (SovCISS), introduced in (Holló et. al 2012). This stress index aims to measure the level of frictions and strains in the financial system of a country and thus to condense that state of financial instability into a single statistic. The underlying data is sourced from the ECB Statistical Data Warehouse, currently available as monthly data for 17 European countries.

3.2 Macro-Financial Indicators

We propose the usage of a number of country-level vulnerability indicators, which are selected to cover a range of macro-financial imbalances. Measures included cover asset prices (e.g., house and stock prices), leverage (e.g., mortgages, private loans and household loans), business cycle indicators (GDP and inflation), measures from the EU Macroeconomic Imbalance Procedure (e.g., current account deficits and government debt), and the banking sector (e.g., loans to deposits). Most indicators use common transformations, such as ratios to GDP or income, growth rates, and absolute and relative deviations from a trend. The data sources are Eurostat, OECD, ECB Statistical Data Warehouse and the BIS Statistics. As the time unit of the indicator data is quarterly, the stress index will require averaging by the mean of three months to obtain quarterly averages applicable for regression.

3.3 Forecasting Methods

Following the notion that a universally applicable single method for any scenario generally cannot be found, the following array of methods is proposed for forecasting. Several of these methods have shown good performance in the classification case with similar data and specifications.

Ordinary Least Squares (OLS). A well-known multivariate linear regression method (see e.g. Plackett (1972)), which minimizes the differences between the predicted values and the actual values.

k-Nearest Neighbors (KNN). A non-parametric method which uses similarity functions to calculate output based on the *k* closest neighbors in the data, see Altman (1992).

Classification and Regression Trees (CART). The regression tree (Breiman et al. 1984) implements a decision tree-type structure, which provides an output by performing a sequence of tests on the predictor values. In relation to the other proposed methods, it is worth mentioning that a regression tree often has a somewhat jagged output due to its construction, particularly for trees of low complexity. The obvious advantage of the method is high interpretability.

Random Forest (RF). As the name suggests, this method grows a "forest" by constructing a number of regression trees (see Breiman (2001)). Randomness in the trees is induced by using differently sampled subsets of the data for each tree, along with considering only a randomly drawn subsample of predictors for each branch. When the average of all trees is calculated, variance is reduced as there is less correlation between the trees.

Support Vector Machine (SVM). Introduced by Cortes and Vapnik (1995), the SVM is one of the most popular machine learning methods for supervised learning. The SVM utilizes hyperplanes in a high-dimensional space to construct a maximum margin separator. In the regression case, a loss function for this separator is implemented, which ignores deviations from the true value within a set tolerance. This feature, in combination with the usage of support vectors (i.e. not all data points are considered), allows the SVM to learn complex non-linear relationships, but still retain generalization ability.

3.4 From Regression Back to Classification

Despite that regression makes better use of the data, as it preserves the variation rather than simplifying output to binary events, policymakers still mostly act in a binary manner. The general decision whether to impose a macroprudential tool or not, while also requiring calibration, is a binary choice. We thus propose to apply simple thresholds to both the regressed variable and the estimated variable to assess the degree to which our models perform in standard classification, or signaling.

4 CONCLUSION

In this short paper, we have outlined the opportunities and challenges in performing early-warning regression of a financial stress index, and put forward a suitable set-up as a modeling framework. Despite not being well-documented in the literature, early-warning regression has an obvious advantage due to it forgoing binary (pre-)crisis events in favor of predicting a continuous stress indicator. The ability of an early-warning model to account for the magnitude of systemic risk, in combination with the

recursive evaluation exercise acknowledging cross-sectional and time-dependence, shows clear potential and calls for closer examination.

REFERENCES

Altman, N. S. 1992, An introduction to kernel and nearest-neighbor nonparametric regression, in: *The American Statistician*, Vol. 46, No. 3, pp. 175-185,

Arlot, S & Celisse, A. 2010, A survey of cross-validation procedures for model selection, in: *Statistical Surveys*, Vol. 4, pp. 40-79.

Breiman. L. 2001, Random forests, in: *Machine Learning*, Vol. 45, No. 1, pp. 5-32.

Breiman, L., Friedman, J. H., Olshen, R. H., & Stone, C. J. 1984, *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Cortes, C. & Vapnik, V. 1995, Support-vector networks, in: *Machine Learning*, Vol. 20, No. 3, pp. 273-297.

Holló, D., Kremer M., & Lo Duca, M. 2012, CISS - A Composite Indicator of Systemic Stress in the Financial System. ECB Working Paper No. 1426.

Holopainen M. & Sarlin, P. 2015, Toward robust early-warning models: A horse race, ensembles and model uncertainty, arXiv:1501.04682.

Plackett, R. L. 1972, Studies in the History of Probability and Statistics XXIX - The discovery of the method of least squares, in: *Biometrika*, Vol. 59, No. 2, pp. 239-251.

Stone, M. 1977, Asymptotics for and against cross-validation, in: *Biometrika*, Vol 64, pp. 29-35.

Challenges and Opportunities of Internet of Things in Healthcare*

Göran Pulkkis[i], Magnus Westerlund[ii], Jonny Karlsson[iii], Jonas Tana[iv]

Abstract

The significance of Internet of Things (IoT) in healthcare is point out. Relevant research questions are discussed. A survey of IoT I healthcare includes a classification of healthcare IoT applications, the concept of IoT in closed loop healthcare services, and presentation of four implemented healthcare IoT applications. Proposals about architecture and data analysis structuring for IoT healthcare applications are presented. Examples of threats against IoT security, the privacy of IoT healthcare users, and IoT healthcare application reliability are given. Proposed IoT security requirements and a proposed vision for IoT in healthcare are described.

Keywords: IoT, healthcare devices, healthcare services, Internet security, network privacy, device reliability, software reliability

1 INTRODUCTION

Healthcare is one of the fastest industries to adopt to Internet of Things (IoT) since integrating IoT features into medical devices improves the quality and effectiveness of service which brings an especially high value for elderly patients with chronic conditions and patients requiring constant monitoring and supervision. IoT devices enable remote health monitoring and emergency notification systems. Examples of health monitoring devices are blood pressure and heart rate monitors, advanced devices for monitoring

* Funding from TUF is gratefully acknowledged.
[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [goran.pulkkis@arcada.fi]
[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [magnus.westerlund@arcada.fi]
[iii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [jonny.karlsson@arcada.fi]
[iv] Arcada University of Applied Sciences, Finland, Department of Health and Welfare, [jonas.tana@arcada.fi]

specialized implants such as pacemakers, electronic wristbands, and advanced hearing aids.

This working paper presents an overview of IoT in healthcare. The purpose is to discuss relevant research questions such as architecture of IoT in healthcare, analysis of data from IoT in healthcare, and development methodologies for IoT applications in healthcare.

2 IOT APPLICATIONS IN HEALTHCARE

Many applications of IoT in healthcare have been proposed, developed, implemented, deployed, and reported. This section presents a proposed classification of IoT applications, a proposed concept of IoT in closed loop healthcare services, three applications described in a recent IoT survey (Al-Fuqaha et al. 2015):

- a nursing home patient monitoring system,
- a system for the monitoring and mitigation of eating disorders,
- an indoor navigation system for blind and visually impaired persons,

and an application for diabetes patients.

2.1 A Classification of IoT Applications in Healthcare

Applications of IoT to healthcare can be grouped into

- tracking of objects and persons
- identification and authentication of persons
- automatic data collection
- sensing.

Tracking includes both real-time position tracking, tracking of motion through choke points, such as access to designated areas, continuous inventory location tracking (for example for maintenance, for availability, and for monitoring of use), and materials/objects tracking for example to prevent left-ins during surgery. Identification and authentication includes patient identification to avoid harmful incidents such as wrong drug/dose/time/procedure, electronic medical record maintenance both in the in- and out-patient settings, and identification of infants in hospitals to prevent mismatching. Automatic data collection includes automated care and procedure auditing as well as medical inventory management. Sensor devices enable remote diagnoses of patient conditions by provision of real-time information on patient health indicators. (Atzori, Iera, & Morabito 2010)

2.2 IoT in Closed Loop Healthcare Services

Dohr et al. (2010) describe conceptually IoT in closed loop healthcare services, where patients are monitored by IoT devices sending collected data to and receiving control from a tele-monitoring service center. Both patients and healthcare personnel (physicians and nurses in hospitals and healthcare stations) have secure access to the service center. Direct communication by mobile phones between health care personnel and pa-

tients, such as medical treatment advice to patients, is based on presentation of collected data in the service center.

2.3 Nursing Home Patient Monitoring System

Patients' vital sign measurements are collected and delivered to multiple nursing stations. Light sensors and door sensors are deployed to monitor the activity level of patients and potentially identify patients suffering from depression with the assumption that patients have private rooms. A sensor is USB connected to a microcontroller programmed as a client of the application level Message Queue Telemetry Transport (MQTT) protocol (MQTT 2014) built on the top of the TCP protocol. The MQTT client publishes collected sensor data to a MQTT broker device utilizing a WLAN. MQTT servers connected to the nursing stations subscribe to the MQTT broker to fetch collected data of interest. Doctors can access collected sensor data remotely from a mobile MQTT server application which subscribes to data having the topics of interest. The MQTT broker, which checks authorization of publishers and the subscribers, can be publically exposed on the Internet behind a firewall through a 4G network connection. (Al-Fuqaha et al. 2015)

2.4 Monitoring and Mitigation of Eating Disorders

The application allows patients with essential tremors or Parkinson disease to eat without spilling food. A glove can equipped with tiny vibrating motors counteracts hand movement instability measured by accelerometers. The accelerometer sensors and vibrating motors must communicate with minimum possible delay to deliver the required functionality. Therefore, the Data Distribution Service (DDS) protocol, a broker-less publish-subscribe protocol for real-time M2M Communications (Data 2016), is the right choice for minimal direct communication between the accelerometers and the vibrating motors without a MQTT broker's involvement. In order to integrate this functionality with the nursing stations, a gateway must be deployed to translate DDS messages to MQTT publish messages sent with WiFi to a MQTT broker device. MQTT servers connected to the nursing stations subscribe to the MQTT broker to fetch collected data of interest. Doctors can access collected sensor data remotely with a mobile MQTT server application as in the Home Patient Monitoring System application. (Al-Fuqaha et al. 2015)

2.5 In-Door Navigation System for Blind and Visually Impaired Persons

A constellation of transceivers provides Real-time locating services (RTLS) to users. In this case, a user held device utilizes the multicast DNS (mDNS) protocol (Cheshire & Krochmal 2013) to connect to a local RTLS server and obtain an authentication token to access RTLS services. The RTLS nodes themselves utilize DDS for timely exchange of data packets. The RTLS nodes can relay their collected data to the local RTLS server in

order for it to estimate the current location of the user. The local RTLS server can overlay the location on a floorplan obtained from an Internet connected server to provide tactile navigation information to the users allowing them to avoid obstacles and other physical movement constrains reported by earlier users of the system. (Al-Fuqaha et al. 2015)

2.6 Insulin Dosage Administration for Diabetes Patients

Diabetes patients can benefit from GoCap, an IoT device developed by a company GoCap is a replacement cap for prefilled insulin pens that records the amount of insulin administered daily and the specific times the dosages were administrated. The recorded information is transmitted using Bluetooth to a mobile phone or to connected glucometer. The purpose is to make a steady stream of relevant information transmitted in an easy format available to healthcare professionals for identification and handling of potential problems early enough before the problems become so severe that the patients require hospitalization. (Baum 2013)

3 ARCHITECTURE FOR IOT IN HEALTHCARE

Pang (2013) presents a Wireless Sensor Network (WSN) architecture and Hospital Information System based architecture for IoT in healthcare. The WSM architecture is a "three-tiered" mobile architecture. The lowest tier is composed of randomly deployed IoT devices called sensor nodes ad hoc networked with Master Sensor Nodes (MSN). MSNs are Wide Area Network (WAN) connected to a Central Server functioning as an Internet gateway. Users with mobile phones, laptops, etc. in the medium tier can move to anywhere at any time with WAN access to WSNs and WAN or Internet access to the Central Server. The highest tier is the number of fixed installed access points (to WLANs, cellular networks base stations in cellular networks, etc.). In the Hospital Information System based architecture there is

- a patient layer with an In-home Healthcare Station, which has
    - a wireless Sensor Area Network connection to wearable sensor and other healthcare IoT devices
    - a network interface to the Hospital Information System
    - user interfaces, for example touch screen based, for the patient and for healthcare personnel
- as a hospital layer an Information System
    - for department and labs of the hospital
    - with a database for Electronic Health Records of patients
    - with Emergency Services
    - network connections to In-Home Healthcare Stations of patients
    - an interface to a Cooperative Health Cloud
- as an ecosystem layer a Cooperative Health Cloud with
    - a Service Repository
    - interfaces to Information Systems of Hospitals, Elderly Houses, Healthcare Means Suppliers, and Public Authorities

4 DATA ANALYTICS FOR IOT IN HEALTHCARE

IoT healthcare applications involve collection, storage and analysis of data received from different types of electronic devices and sensors. The amount of generated data is too large for medical staff to analyze manually and therefore automated systems must be used for this purpose. Analytical models for automatic health monitoring and medical diagnosis making based on collected data from sensors are needed. (Nambiar et al. 2013)

Tyagi, Darwish, & Khan (2014) propose a layered framework for management of IoT data. This framework is also applicable to IoT applications in healthcare. Following 6 layers are proposed for IoT data transformations:

1. *Raw data* from IoT devices as data sources
2. *Non-structured data* with data collection and registration
3. *Classified structured data* with data processing such as filtering and enriching
4. *Processed data* with data analysis such as classification and prediction
5. *Viewable data* with delivery and visualization
6. *Information* for clients.

These IoT data transformation layers are related to 7 IoT data management layers:

1. Data acquisition/production (RFID tags, sensors etc.)
2. Data pre-processing (cleaning, removal of duplicates)
3. Data storage/management and archival
4. Specialized data storage (Data mining techniques, partitioning. Data curation techniques)
5. Data analysis and real time event processing
6. Visualization techniques
7. Application

5 SECURITY, PRIVACY, AND RELIABILITY ISSUES

Prevention of, mitigation of, and other defense against several possible security attacks and threats is a necessity for all IoT applications in healthcare. Also the privacy of healthcare IoT application users and the operational reliability of healthcare IoT applications must be granted.

5.1 Threats

Examples of security threats related to IoT in healthcare in (Security 2015):

- The body of a person can be accessed and manipulated causing injury or worse through unauthorized access to physical sensing, actuation and control systems such as implantable and non-implanted medical devices
- Health care providers can improperly diagnose and treat patients based on modified health information or manipulated sensor data
- Malicious parties can steal identities and money based on leakage of sensitive information such as Personal Health Information (PHI)

- Leakage of personal health information can occur by aggregating data from many different systems and sensors
- Unauthorized tracking of people's locations can occur through usage pattern tracking based on healthcare asset usage time and duration
- Unauthorized tracking of a person's behavior and activities can occur through examination of location-based sensing data that exposes patterns and allows analysis of activities, often collected without explicit notice to the target person
- Unlawful surveillance through persistent remote monitoring capabilities offered by small-scale IoT devices
- Inappropriate profiles and categorizations of individuals can be created through examination of network and geographic tracking and IoT metadata
- Vandalism, theft or destruction of IoT assets that are deployed in remote locations and lack physical security controls
- Ability to gain unauthorized access to IoT edge devices to manipulate data by taking advantage of the challenges related to updating software and firmware of embedded healthcare devices
- Ability to create botnets by compromising large quantities of IoT edge devices
- Ability to impersonate IoT devices by gaining access to keying material held in devices that rely upon software-based trust stores.

5.2 Security Requirements

Following standard security requirements for IoT systems are listed in (Tarouco et al. 2012):

- Resilience to attacks – There must be no single points of failure and the system should adjust itself to node failures;
- Data authentication – Authentication of retrieved addresses and object information is a necessity;
- Access control - Information providers must implement access control on provided data;
- Client privacy - Only the information provider should be able to infer from observing the use of the lookup system.

6 FUTURE OUTLOOK

Kramp, van Kranenburg, & Lange (2013) present following vision for IoT in healthcare:

> "*Control and prevention are two of the main goals of future health care. Already today, people have the option of being tracked and monitored by specialists even if the patient and specialist are not in the same place. Tracing peoples' health history is another aspect that makes IoT-assisted e-health very versatile. Business applications could offer the possibility of medical services not only to patients but also to specialists, who need information to proceed in their medical evaluation. In this domain, IoT makes human interaction much more efficient because it permits not only localization, but also tracking and monitoring of patients.*"

7 CONCLUSIONS

In the development of healthcare IoT applications cooperation between computer scientists, ICT professionals, healthcare professionals, healthcare researcher, and users of these applications is a necessity. Interoperability between healthcare IoT devices and software component emphasizes widely accepted IoT architecture standards. Analysis of IoT data is a big data issue already for a single user of IoT healthcare application and particularly for the whole user population. Healthcare IoT applications must in addition to IoT security requirement also protect the privacy of users and provide practically fault-free reliability to safeguard the life and safety of users.

REFERENCES

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. 2015, Internet of Things: A Survey on EnablingTechnologies, Protocols, and Applications, in: *IEEE Communication Surveys & Tutorials*, Vol. 17, No. 4, pp. 2347-2376

Atzori, L., Iera, A., & Morabito, G. 2010, The Internet of Things: A survey, in: *Computer Networks*, Vol. 54, Iss. 15, pp. 2787-2805

Baum, S. 2013, A remote monitor embedded in insulin pen caps could help personalize diabetes treatment. MedCityNews. Accessed 2.6.2016. Published 3.6.2013. http://medcitynews.com/2013/06/a-remote-monitor-embedded-in-insulin-pen-caps-could-help-personalize-diabetes-treatment/?rf=1

S. Cheshire, S. & Krochmal, M. 2013, Multicast DNS, RFC 6762, Internet Engineering Task Force (IETF).

Data Distribution Service™ (DDS™) 2016 Accessed 5.6.2016. Published 2016. http://www.omg.org/spec/DDS/

Dohr, A., Modre-Opsrian, R., Drobics, M., Hayn, D., & Schreier, G. 2010, The Internet of Things for Ambient Assisted Living, in: *Proceedings of the 2010 Seventh International Conference on Information Technology: New Generations*, IEEE Press, pp. 804-809.

Kramp, T., van Kranenburg, R., & Lang, S. 2013, Introduction to the Internet of Things, in: Bassi, A., Bauer, M., Fiedler, M., Kramp, T., van Kranenburg, R., Sebastian Lange, S., & Meissner, S. (Eds.) 2013, *Enabling Things to Talk. Designing IoT solutions with the IoT Architectural Reference Model*, Springer, Heidelberg. http://dx.doi.org/10.1007/978-3-642-40403-0

MQTT 2014, Accessed 4.6.2016. Published 7.11.2014. http://mqtt.org/

Nambiar, R., Sethi, A., Bhardwaj, R., & Vargheese, R. 2013, A Look at Challenges and Opportunities of Big Data Analytics in Healthcare, in: *Proceedings of the 2013 IEEE International Conference on Big Data*, IEEE Press, pp. 17-22.

Pang, Z. 2013, Technologies and Architectures of the Internet-of-Things (IoT) for Health and Well-being, Doctoral Thesis in Electronic and Computer Systems, KTH – Royal Institute of Technology, Stockholm, Sweden, Accessed 4.6.2016. Published January 2013. http://kth.diva-portal.org/smash/get/diva2:621384/FULLTEXT01.pdf

Security Guidance for Early Adopters of the Internet of Things (IoT) 2015, CSA Cloud Security Alliance, Mobile Working Group. Accessed 2.6.2016. Published April 2015. https://downloads.cloudsecurityalliance.org/whitepapers/Security_Guidance_for_Early_Adopters_of_the_Internet_of_Things.pdf

Tarouco, L.M.R., Bertholdo, L.M., Granville, L.Z., Arbiza, L.M.R., Carbone, F., Marotta, M., & de Santanna, J.J.C. 2012, Internet of Things in Healthcare : Interoperability and Security Issues, in: *Proceedings of the 2012 IEEE International Conference on Communications (ICC)*, IEEE Press, pp. 6121-6125.

Tyagi, S., Darfish, A., & Khan, M.A. 2014, Managing Computing Infrastructure for IoT Data, in: *Advances in Internet of Things*, Vol 4, pp. 29-35. http://dx.doi.org/10.4236/ait.2014.43005