Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA 19102

Grantee name: Barbara Lewandowska-Tomaszczyk

Details of the STSM

Title: **BIAS CONCEPTUALIZATION, IDENTIFICATION AND MODERATION IN ONLINE MEDIA**

Start and end date: 19/06/2022 to 25/06/2022

Description of the work carried out during the STSM

The aim of this STMS has been a study of language in (social) media and the extent and scope of biases there by means of computational and linguistic methods.

In our daily discussions with the Host Dr Sviatlana Höhn and the team members of the Computer Science Department of the University of Luxemburg, the conceptualization of types of biases in online media content and bias identification approaches have been reviewed. We also made a survey of digital tools for bias identification and discussed its links with the categories of offensive language as well as discourse functions of biased messages in online media and social media. The next of the vital topics connected with this research area is the need, necessity, and possibilities of strict bias control in the internet content, as well as approaches to content moderation. That's why reference to the identification and prediction of offensive content in social media has been made and examples of the link between such harmful content and internet interactants' reciprocal offence threats pointed out.

Discussions on the planned joint conference presentations and publications have taken place on regular basis, aiming at the identification and interpretation of bias distinguishing types and its manifestation in various domains, with some proposals of moderating intervention after harmful biases are identified in online content posted by the media and individual users. We analysed over sixty newspaper articles in four different languages: Russian, Polish, English, and German

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

to prepared the date for the forthcoming presentation and papers. considered a comparative material in the analysis.

The meetings at the Department were held between the applicant and the Host Dr Sviatlana Höhn, as well as with the team members: Dr Sjouke Mauw and other colleagues, following a talk I was asked to give on *Offense, Emotion and Biased Persuasion in the Media.*

(max. 500 words)

Grantee enters max 500 word summary here.

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

(max. 500 words)

As the main theme of the STSM is bias conceptualization in online media, online media content and bias typology, bias identification approaches and digital tools have been reviewed and the methodologies, analysis and harmful content de-biasing approaches proposed.

We plan the following presentations and publications (titles for Paper 2 and Paper 3 are provisional).

Presentation 1.
A number of online articles on a tragic fire event in 2014 in Odessa (involving pro-unity and pro-federation activists in Ukraine) were analysed to be discussed and presented at the coming LITHME conference in Jyväskylä from the semantic and computational perspectives. English, Russian, Polish and German texts are a comparative material in this presentation, which concludes with a discussion of a proposed model of de-biasing methodology.

Presentation at the 2nd COST Action LITHME meeting at the University of Jyväskylä (30 June-1 July 2022).

Understanding Content Bias: qualitative research supports computational methods

Sviatlana Höhn, Barbara Lewandowska-Tomaszczyk, Dov Gabbay and Sjouke Mauw
Linguistically biased news articles are flooding the Internet. Disinformation campaigns are not simply based on lies, but build on biased public opinions and stereotypes that populations believe in. Multiple computational approaches have been developed to detect and filter disinformation, misinformation, and biased news. Several attempts have been proposed to correct bias in texts. The approaches vary from wordlist based over similarity metrics and vectorized representations of language, to complex machine-learning architectures. Recent critical reviews point to the lack of proper understanding and definitions of bias in most

approaches. Therefore, it becomes difficult to evaluate their actual usefulness. In our talk, we use Membership Categorization Analysis methods to demonstrate how biased news text can be created and how lies are made believable in disinformation campaigns targeting particular populations and (mis)using their beliefs. we introduce a new implementable model for linguistic bias research. The model includes simple operations that can be used for biasing and de-biasing of text for a particular audience. Finally, we will explain which types of language resources are needed for the model to work.

Presentation and paper 2.

Analysing Bias: taxonomy, structure, and de-biasing approaches

Sviatlana Höhn, Barbara Lewandowska-Tomaszczyk, Dov Gabbay and Sjouke Mauw

Presentation at:

The Thirty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-23) *Collocated with AAAI-23* | February 7-14, 2023, Washington, DC, USA

Paper submission by 8 November 2022.

Paper 3.

Bias, Offense and Emotionality in the (Social) Media

Barbara Lewandowska-Tomaszczyk and Sviatlana Höhn

The paper explores links between the definition and typologies of bias events, their links with the idea of explicit and implicit offensive effect and biased persuasive emotionality present in this communication types. Examples of biased discourses and their language- and culture-specific character are provided and their more rigorous analysis fr computational applications proposed.

To be submitted to *Lodz Papers in Pragmatics* (Mouton de Gruyter) by 15 January 2023.

Grantee enters max 500 word summary here.