Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 101-104 recibido 02-04-2020 revisado 24-05-2020 aceptado 24-05-2020

MISMIS: Misinformation and Miscommunication in social media: aggregating information and analysing language

MISMIS: Desinformación y agresividad en los medios de comunicación social: agregando información y analizando el lenguaje

Paolo Rosso[1], Francisco Casacuberta[1], Julio Gonzalo[2], Laura Plaza[2], J. Carrillo[2], E. Amigó[2], M. Felisa Verdejo[2], Mariona Taulé[3], Maria Salamó[3], M. Antònia Martí[3]
[1] PRHLT-Universidad Politécnica de Valencia
[2] NLP&IR- UNED
[3] CLiC-Universitat de Barcelona
{prosso,fcn}@dsic.upv.es; {julio, lplaza, jcalbornoz, enrique, felisa}@lsi.uned.es;
{mtaule, maria.salamo, amarti}@ub.edu

Abstract: The general objectives of the project are to address and monitor misinformation (biased and fake news) and miscommunication (aggressive language and hate speech) in social media, as well as to establish a high quality methodological standard for the whole research community (i) by developing rich annotated datasets, a data repository and online evaluation services; (ii) by proposing suitable evaluation metrics; and (iii) by organizing evaluation campaigns to foster research on the above issues.
Keywords: Fake news, biased information, hate speech, social media

Resumen: Los objetivos generales del proyecto son abordar y monitorizar la desinformación (noticias sesgadas y falsas) y la mala comunicación (lenguaje agresivo y mensajes de odio) en los medios de comunicación social, así como establecer un estándar metodológico de calidad para toda la comunidad investigadora mediante: i) el desarrollo de *datasets* anotados, un repositorio de datos y servicios de evaluación online; ii) la propuesta de métricas de evaluación adecuadas; y iii) la organización de campañas de evaluación para fomentar la investigación sobre las cuestiones mencionadas.
Palabras clave: Noticias falsas, información sesgada, mensajes de odio, medios de comunicación social

1 Introduction

Social media have become the default channel for people to access information and express ideas and opinions. The most notable and positive effect is the democratization of information and knowledge and the capacity of influence on the public opinion. Instead of a few media that control which information spreads and how (radio and TV channels, news media, etc.), any citizen can now share her views with the world and, potentially, influence the public state of opinion on any topic. But there are also undesired effects of this democratization of knowledge that is becoming increasingly important. One of them is that social networks foster information bubbles: every user may end up receiving only the information that matches her personal biases, beliefs, tastes and viewpoints. A perverse effect is that social networks are a breeding ground for the propagation of fake news: when a piece of news outrages us or matches our beliefs, we tend to share it without checking its veracity; and, on the other hand, content selection algorithms in social networks give credit to this type of popularity because of the click-based economy on which their business are based. Another harmful effect is that the relative anonymity of social networks facilitates the propagation of toxic, hate and exclusion messages. Therefore, social networks contribute, paradoxically, to misinformation and miscommunication. The main aim of this project is to alleviate these perverse effects via language analysis and the computational modelling of the propagation mechanisms for factual information, ideas and opinions in social networks.

ISSN 1135-5948. DOI 10.26342/2020-65-13

1.1 Participants

The MISMIS project (PGC2018-096212-B) is coordinated by Paolo Rosso (Universidad Politécnica de Valencia, UPV) and funded by the Spanish Ministry of Science, Innovation and Universities (R+D Knowledge Generation program). It started in 1st January 2019 and it will run until 31st December 2021. These are the groups involved in this project:

- The PRHLT[1] research center at UPV, which leads the MISMIS-FAKEnHATE (PGC2018-096212-B-C31) subproject. This is an experienced group in pattern recognition, machine learning, multimodality and NLP (e.g. author profiling, stance detection and automatic misogyny identification).
- The NLP&IR[2] research group at Universidad Nacional de Educación a Distancia (UNED), which leads the MISMIS-BIAS (PGC2018-096212-B-C32) subproject. They contribute to the project with their experience in the area of evaluation metrics and methodologies for Information Retrieval and NLP.
- The CLiC[3] research group at Universitat de Barcelona (UB), which leads the MISMIS-LANGUAGE (PGC2018-096212-B-C33) subproject. This group will contribute to this project with the development of theoretically founded linguistic models for the detection of misinformation and miscommunication, as well as by the creation of the datasets necessary for training and evaluating the systems that will be carried out.

2 Hypotheses and Objectives

The two main hypotheses that will guide our research are:

H1. Misinformation and miscommunication should be linguistically modelled, not only to improve the performance of algorithms but also to provide a better understanding of these phenomena.

H2. Linguistic resources, algorithms and metrics properly developed to address misinformation and miscommunication are essential for a consistent advancement related to the following general objectives of the project.

These objectives are the following:

- **O1 - Misinformation.** We will model and develop techniques to overcome the problem of misinformation from three angles: (i) by developing methods to extract unbiased knowledge from biased sources (what we call the wisdom of biased crowds); (ii) by developing techniques to identify fake news; and (iii) by designing summarization and visualization techniques for controversial topics, in which biases and manipulated information are exposed and the different points of view are identified and summarized contrastively.
- **O2 - Miscommunication.** We will develop techniques to detect and monitor aggressiveness and hate speech in social media, which contribute to create increasingly polarized communities and to demote argumentation in favour of pure confrontation. We will model hate speech from a linguistic point of view in different languages and we will also use multimodal signals (images and video) to complement textual evidence.
- **O3 - Methodological Tools.** One of the distinguishing aspects of our proposal is that we will make a substantial effort to improve research methodologies in the field, with the goal of establishing a high methodological standard for the research community. This is a horizontal objective for which we will: (i) develop rich annotated datasets for each of the relevant tasks; (ii) perform a formal study of suitable evaluation metrics that appropriately define each of the tasks and provide adequate analytical tools for evaluating system performance; (iii) develop a data repository and an online evaluation service in which metrics and datasets are centrally available for the research community, and (iv) organize evaluation campaigns using all the above tools to foster the sharing and reuse of knowledge in the research community.

3 Work progress

The problem of misinformation detection (O1) has been addressed from several perspectives: rumour detection (Ghanem et al., 2019a), fact checking (Ghanem et al., 2019b), fake news detection (Ghanem et al., 2019c) and credibility detection (Giachanou et al., 2019). False claims, and false information in general, are intentionally written to evoke emotions to the readers in an attempt to be believed and be disseminated in social media. To differentiate between credible and non-credible claims in the above latter works, we incorporated emotional

[1] https://www.prhlt.upv.es/wp/
[2] https://sites.google.com/view/nlp-uned/home
[3] http://clic.ub.edu/en

features into a Long Short Term Memory (LSTM) neural network. In order to address fake news from an author profiling perspective, currently we are organizing a shared task on profiling fake news spreaders on Twitter[4] (O3).

Another way to overcome the problem of misinformation is developing methods to extract unbiased knowledge from biased sources. Recent literature on recommender systems has focused on the inherent biases present in the data, which are amplified by recommendation algorithms. We are currently working on real-world datasets in order to characterize the existence of a geographic bias in the data. We are analysing the bias of current state-of-the-art recommender systems and try to understand the visibility that is given to items. Next, we will address the mitigation of the bias in the recommendation algorithms. We are also analysing fairness in session-based recommender systems (Ariza et al., 2020). We study how the effectiveness of state-of-the-art algorithms (neural and non-neural approaches) is affected by specific dataset structure, due to different session lengths, in terms of accuracy and distribution of content provider exposure.

Regarding the problem of miscommunication, and more concretely of hate speech detection, we organised the HatEval shared task at SemEval with Twitter corpora in English and Spanish (O3) and immigrants and women as target of hate speech (Basile et al., 2019). The twin problem of identifying offensive tweets (O2) in general (OffensEval shared task at SemEval) was addressed with a deep-learning ensemble method (De La Peña and Rosso, 2019). Special attention was given to the problem where targets of hate speech were women (Frenda et al., 2019a; 2019b). An extensive study of the different forms in that sexism attitudes and behaviours are found in social media has been performed, and different methods for automatically detecting everyday sexism in online communications have been developed (Rodríguez et al., 2020).

Hate speech in social media contributes to create increasingly polarized communities (Lai et al., 2019; 2020). At the moment, we are investigating the impact of hate speech and polarization (O2) on demoting argumentation[5] when controversial issues are addressed (Esteve et al., 2020). The preliminary work carried out to detect fake news and hate speech was covered also by the press[6].

Regarding the corpora developed (O2 and O3), we have created the NewsCom corpus, which contains 2,955 comments in Spanish posted in response to eighteen different news articles from online newspapers. This corpus has been annotated with the focus of negation, scope and negation markers (Taulé et al., 2020). Currently, we are annotating the comments of this corpus indicating their degree of toxicity: not toxic, midly toxic, toxic, very toxic. We are enriching this annotation with new and more detailed criteria. This annotation allows us to establish fine grained criteria for analysing and better defining what can be considered as a comment with toxic, offensive, abusive or hate language. This corpus could be used for the development of algorithms for the detection of this kind of language. Another line of research is the creation of a corpus of fake news from data provided by the Newtral[7] start-up with the aim to be used for training and testing algorithms for textual-based and multimodal fake news detection. These corpora could be used in future evaluation campaigns.

Regarding our work on methodological foundations of evaluation and textual similarity (O3): (i) we have completed a study on the evaluation of ordinal classification tasks (Amigó et al., 2020). Such tasks are very relevant for NLP and for the project, but they are currently evaluated with metrics for different problems (classification and regression). We have defined formal restrictions for the task and we have proposed a new metric, based on Information Theory and Measurement Theory, which complies with all the formal requisites and behaves better empirically than available alternatives. (ii) We have also completed the first stages of a formal study on similarity functions, which includes a specific axiomatics for similarity in the context of NLP and a new similarity function, which is most robust than state of the art alternatives. (iii) We have addressed the task of semantic compositionality, comparing word2vec approaches with contextual embeddings and

[4] https://pan.webis.de/clef20/pan20-web/author-profiling.html

[5] http://www.argumentsearch.com/

[6] https://www.lavanguardia.com/tecnologia/20191207/472082286311/usan-la-ia-para-detectar-noticias-falsas-y-mensajes-de-odio-en-redes-sociales.html

[7] https://www.newtral.es/

exploring unsupervised alternatives with a formal foundation.

Acknowledgments

The MISMIS project (PGC2018-096212-B) is funded by the Spanish Ministry of Science, Innovation and Universities.

References

Amigó, A., J. Gonzalo, S. Mizzaro and J. Carrillo. 2020. Effectiveness Metrics for Ordinal Classification: Formal Properties and Experimental Results. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

Ariza, A., F. Fabbri, L. Boratto and M. Salamó. 2020. From The Beatles to Billie Eilish: Connecting Provider Representativeness and Exposure in Session-based Recommender Systems. Submitted to *SIGIR 2020*.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel, P. Rosso and M. Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *Proc. of the 13th Int. Workshop on Semantic Evaluation* (SemEval-2019), (NAACL-HLT 2019), Minnesota, USA, 54-63.

De La Peña, G. P. and P. Rosso. 2019. DeepAnalyzer at SemEval-2019 Task 6: A Deep Learning-based Ensemble Method for Identifying Offensive Tweets. *Proc. of the 13th Int. Workshop on Semantic Evaluation* (SemEval-2019), NAACL-HLT 2019, Minnesota, USA, 582–586.

Esteve, M., F. Casacuberta and P. Rosso. 2020. Minería de argumentación en el Referéndum del 1 de Octubre de 2017. *Procesamiento del Lenguaje Natural*, 65. To appear.

Frenda S., N. Kando, V. Patti and P. Rosso. 2019a. Stance or Insults?. *Proc. of the Ninth International Workshop on Evaluating Information Access* (EVIA 2019), Satellite Workshop of the NTCIR-14 Conference, National Institute of Informatics, Tokyo, Japan.

Frenda, S., B. Ghanem, M. Montes-y-Gómez and P. Rosso. 2019b. Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5): 4743–4752.

Ghanem B., A. Cignarella, C. Bosco, P. Rosso, and F. Rangel. 2019a. UPV-28-UNITO at SemEval-2019 Task 7 Exploiting Post's Nesting and Syntax Information for Rumor Stance Classification. *Proc. of the 13th Int. Workshop on Semantic Evaluation* (SemEval-2019), NAACL-HLT 2019, Minnesota, USA, 1125–1131.

Ghanem, B., G. Glavas, A. Giachanou, S. Ponzetto, P. Rosso and F. Rangel. 2019b. UPV-UMA at CheckThat! Lab: Verifying Arabic Claims using Crosslingual Approach. L. Cappellato, N. Ferro, D. E. Losada and H. Müller (eds.) *CLEF 2019 Labs and Workshops, Notebook Papers*. Workshop Proceedings.CEUR-WS.org, vol. 2380.

Ghanem B., P. Rosso and F. Rangel. 2019c. An Emotional Analysis of False Information in Social Media and News Articles. (arXiv:1908.09951). *ACM Transactions on Internet Technology* (TOIT). To appear.

Giachanou A., P. Rosso and F. Crestani. 2019. Leveraging Emotional Signals for Credibility Detection. *Proc. of the 42nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '19), Paris, France.

Lai, M., M. Tambuscio, V. Patti, G. Ruffo and P. Rosso. 2019. Stance Polarity in Political Debates: a Diachronic Perspective of Network Homophily and Conversations on Twitter. *Data & Knowledge Engineering*, 124. https://doi.org/10.1016/j.datak.2019.101738

Lai, M., V. Patti, G. Ruffo and P. Rosso. 2020. #Brexit: Leave or Remain? The Role of User's Community and Diachronic Evolution on Stance Detection. *Journal of Intelligent & Fuzzy Systems*. To appear.

Rodríguez-Sánchez, F. 2019. *Desarrollo de un sistema para la detección del machismo en redes sociales.* Trabajo Final de Máster en Tecnologías del Lenguaje. UNED.

Taulé, M., M. Nofre, M. González and M.A. Martí. 2020. Focus of negation: its identification in Spanish. *Natural Language Engineering*,1-22. https://doi.org/10.1017/S1351324920000388