AI for Hate Speech Detection in Social Media*

Andres Montoro[1], Jose A. Olivas[1][0000-0003-4172-4729] and Adan Nieto[2][0000-0002-7899-4725]

[1] Department of Information Technologies and Systems, University of Castilla-La Mancha, 13071 Ciudad Real, Spain
andres.montoro@alu.uclm.es, joseangel.olivas@uclm.es
[2] Department of Public and Corporate Law, University of Castilla-La Mancha, 13071 Ciudad Real, Spain
adan.nieto@uclm.es

Abstract. The main goal of this work focuses on solving the problem of analyzing the data coming from Social Media and exploring the mechanisms for the extraction and representation of knowledge from all the different disciplines outside the world of Information Technologies. Soft Computing and Big Data techniques are used to deal with the challenges mentioned. This paper shows a mechanism to detect hate speech in Social Media using Soft Computing and Sentiment Analysis, and it also stablishes the base of a doctoral thesis.

Keywords: Soft Computing, Fuzzy Logic, Computing with words, Social Media Mining, Big Data, Text Mining.

1 Introduction.

In the age of Big Data millions of people are generating data in social media. This kind of data is unstructured, noisy, non-formatted and has variable length. Inside the universe of social media, the relations between entities (Social networks) becoming an extraordinary vehicle for the bulk dissemination of messages.

Effective social media analysis requires collecting information about individuals (users) and entities (social networks, sites, etc), analyzing the interactions between them and discovering patterns to understand human behavior [10].

The analysis of social media presents many challenges that basic techniques of Natural Language Processing (NLP) or Text Mining [1] cannot resolve. Some of these challenges are the Big Data Paradox [10], in other words, the volume of data from social media analysis is clearly Big, which also implies the problem of analyzing data in real time. Noise Removal Fallacy [10], social media data has a lot of noise and a blind removal of this can cause loss of knowledge and the definition of noise could change

* This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grants MERINET and SAFER: TIN2016-76843-C4-2-R and PID2019-104735RB-C42 (AEI/FEDER, UE).

according to the problem to be solved. Cross media data [9] i.e. how to exploit diverse data coming from social media (text, links, multilingual data, slang text, and so on).

2 Motivation.

Words play the main role in social media analysis and overall in human information processing. When we work with words we struggle with imprecision. The concept of computing with words was develop by Zadeh in [7]. In short, it's a field closely related with Fuzzy logic and Soft Computing in which the items to be computed are words, phrases and propositions drawn from a natural language [8].

Soft Computing is born as a set of techniques that groups the use of fuzzy methodologies. It was defined in 1994 by Zadeh [6] as a mixture of different methods that in one way or another cooperate from their foundations. The main components of Soft Computing are: Fuzzy Logic, Probabilistic Reasoning systems, Neural Networks, and either Evolutionary computing [2] or Metaheuristics [5].

Our investigation is focused in this field, using soft computing to deal with the problems found in social media analysis and extending the application of this field to other human disciplines like law or criminology.

3 Case study: Sentiment Analysis for the prevention of hate speech in social media.

The Internet has changed the conditions of communication in society and has become a new area of criminal opportunity different from that of the physical world [3] due to, among other things, its characteristics of neutrality, absence of censure and its constant development. This has led to a wider dissemination of hate crimes and therefore a greater effect.

In this case, it is developed a computational mechanism capable of identifying and classifying according to their intensity, hate messages in social media using techniques of Sentiment Analysis, Natural Language Processing and Fuzzy Logic. The starting point is a taxonomy designed from the current legality and the knowledge of an expert allows to determine the intensity of the hate speech and the particularities that compose it to inform of the pertinent decisions to be taken considering the prevalent legality and the corporate social responsibility of each company.

In literature there exist many approaches to identifying hate speech, in [4] resumes many of that using Natural Language Processing and Machine Learning. Some of its mentioned approaches are:

- Based in message characteristics.
- Using corpora to detect hate terms.
- Meta-information to encourage the model.
- Classification methods.
- Sentiment analysis.

3.1 Model.

All the previous approaches show the diversity of hate speech detection methodologies on the web.

The study not only identifies hate speech, it is also able to classify each message according to its intensity using knowledge engineering and soft computing. This process was developed following these phases:

1. Knowledge acquisition establishes the basis for the development of the taxonomy for the identification of Violent and Hateful Comments.
2. Extraction of ontology from the domain assists in the extraction of hate terms resulting from a message gathering experiment.
3. Taxonomy of violent and hateful communication. It is the result of modelling all the knowledge extracted in the form of a knowledge map.
4. Detection of violent and hateful communication using natural language processing techniques and the ontology.
5. Construction of fuzzy models based on the designed taxonomy, making use of linguistic labels extracted from the dataset using sentiment analysis techniques.

The result is a computer system capable of identifying and classifying hate messages in social media.

3.2 Prototype

Despite the knowledge extracted from the expert, the main source of knowledge is the article 510 of the Spanish Criminal Code. The interpretative framework is very wide, and the legally protected right refers to multiple groups. For the development of a functional prototype, the target group of hatred has been established to be Arabs and/or Muslims.

The prototype has been developed to detect hate messages in Spanish given the criminal awareness component of the model based on Spanish legislation. Its main design is based on the architecture of a fuzzy rule-based system adapted to the domain of the problem. This system is composed by the following steps:

- Input:
  - Gathering potential hate message.
- Fuzzification interface:
  - Obtaining values from ontology using natural language processing to extract relevant terms in the message.
- Inference mechanism:
  - Using taxonomy to label the extracted atomic expressions.
  - Establishing the linguistic labels of the proper fuzzy model, which is built with the knowledge base obtained with the rules derived from the knowledge extraction mechanism.
- Defuzzification interface:
  - Grouping values and extract membership grade.

- Output:
  — Reporting the intensity of hate speech and the reasoning flow.

Scaling to other languages would not be limiting. Thanks to the fuzzy system the taxonomy becomes a reasonably accurate metric for measuring the intensity of hate speech.

4 Future Work.

The main purpose of this paper is to show our investigation line and draw attention to Soft Computing and Big Data analytics like emergency fields in the coming years.

The immediate work focuses on improving the treatment of hate speech in social media exposed in the case study, detecting fake news from the perspective of soft computing and introducing the world of law using computational intelligence to model compliance system from organization analysis through risk assessment and prediction to recommendation systems. This is the starting point for the preparation of a doctoral thesis focused in applying Soft Computing and Big Data Analytics techniques to solve problems in different fields. One of them is law thanks to our collaboration with the Institute of European and International Criminal Law[1].

References

1. Aggarwal C.C., Zhai C. An Introduction to Text Mining. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. (2012).
2. Bonissone, P. P. Soft computing: the convergence of emerging reasoning technologies. Soft Computing - A Fusion of Foundations, Methodologies and Applications, 1(1), 6–18. (1997).
3. Miró Llinares, F. El cibercrimen. Fenomenología y criminología de la delincuencia en el ciberespacio (1st ed.). Madrid, España: Marcial Pons (2012).
4. Schmidt, A., & Wiegand, M. A Survey on Hate Speech Detection using Natural Language Processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 1–10. (2017).
5. Verdegay, J. L., Yager, R. R., & Bonissone, P. P. On heuristics as a fundamental constituent of soft computing. Fuzzy Sets and Systems, 159(7), 846–855. (2008).
6. Zadeh, L. A. Fuzzy logic and soft computing: issues, contentions and perspectives (pp. 1–2). Proc. IIZUKA'94: 3rd Int. Conf. on Fuzzy Logic, Neural Nets and Soft Computing, Iizuka, Japan (1994).
7. Zadeh, L. A. Fuzzy logic = computing with words. IEEE Transactions on Fuzzy Systems, 4(2), 103–111. (1996).
8. Zadeh, L. A. Computing with Words: Principal Concepts and Ideas (Studies in Fuzziness and Soft Computing) (1st ed.). Heidelberg, Berlin: Springer (2012).
9. Zafarani, R., & Liu, H. Connecting Corresponding Identities across Communities. International AAAI Conference on Web and Social Media, 41–49. (2009).
10. Zafarani, Reza, Abbasi, M. A., & Liu, H. Social Media Mining. Cambridge University Press, (2009).

[1] https://blog.uclm.es/idp/