Ethical Implications of Analyzing Opinions, Emotions and Interactions in Social Media

Viviana Patti and Rossana Damiano and Cristina Bosco
Dipartimento di Informatica, Università degli Studi di Torino, Italy
Email: {patti,rossana,bosco}@di.unito.it

***Abstract*—The development of Artificial Intelligence techniques for human language processing implies not only new opportunities for research and industry, but also new responsibilities that the NLP and Computational Linguistics community as a whole must carefully take care of. At the same time, the availability of these tools is itself an opportunity for analysing human biases and preventing abuses by social media users. This position paper proposes an holistic reflection on the opportunities and risks brought about analysing human expressions in interaction, with the main aim of highlighting the need to acquire a new awareness of the possible non-ethical uses of automatic human-processing tools and the potential of their ethical uses.**

1. Introduction

Artificial intelligence techniques for human language processing have been remarkably evolving over the last decade, coupled with the availability of large amounts of user–generated data on social media and more powerful computing tools. These techniques play an increasingly relevant role in various application areas, such as commercial and policy making contexts. They can therefore influence the decision processes to which people daily participate as an active or passive part, e.g. in interfaces that interact with the user through voice commands, but also in various types of analysis and human behaviour prediction tools. For example, to detect political orientation or the level of well-being perceived by individuals, automated opinion-mining & sentiment analysis techniques are often applied, especially to daily and spontaneous textual productions of users within social media such as Twitter, Instagram, Facebook, or sites dedicated to the production and distribution of reviews like TripAdvisor and Amazon Review. The same kind of technology is also used to detect terrorist, homophobic, misogynist, or racist content transmitted through the network, or in medical applications to early identify dementia signals in patients at risk of developing them.

On the one hand, social media data offer interesting and valuable exploration opportunities in many areas. On the other hand, data analysis and data aggregation technologies built within Natural Language Processing (NLP), which might appear to be free from ethical implications, can be, instead, a prerequisite for the development of more or less virtuous applications. Therefore, the development of this area, implies not only new opportunities for research and industry, but also new responsibilities and new reflections that the NLP and Computational Linguistics community as a whole must carefully take care of. What social data is ethical to deal with and what is ethical to infer from those data, considering that from digital traces where users express emotions and opinions on almost anything it is possible to extract information about people's personal traits (age, gender, personality) and space-time movements? How to encourage accountability and transparency when dealing with those data? How to manage the presence of human bias, also linked to prejudices and stereotypes, in user-generated texts? This paper proposes a holistic reflection on these issues, with the main aim of highlighting the need to acquire a new virtuous technological and social awareness of the possible non-ethical uses of automatic human-processing tools, with particular emphasis on possible hazards associated with the combined use of analytical techniques, linguistic texts of social media and personal data accompanying them. On the other hand, we intend to reflect on the opportunities for ethical research enabled by social media, which offer a rich set of data and interactions among users from which to identify, monitor and thus counteract mystifying, discriminatory and hostile attitudes, bias, stereotypes and prejudices [1]. These phenomena can be observed "in the wild" in textual expressions and spontaneous dialogues among social media users and can be analyzed in depth through NLP (hate speech detection, deception detection and so on), aiming to provide an ethical contribution in terms of knowledge that can be used to counteract these trends.

Further challenges come from the use of language technologies in the field of virtual agents, increasingly integrated into social media as virtual assistants and chatbots. Developed to create applications that involve the direct interaction with human beings, virtual agents have gone through a remarkable development over the last two decades, also thanks to the evolution of NLP and Natural Language Generation (NLG) techniques, which allow creating artificial agents who converse with humans by supporting them in various tasks. Agents can also be endowed with a graphic design that increases their expressiveness. From chatbots, applications based on spoken or written dialogue, to embodied characters, the field of virtual agents is booming, not without ethical implications. By relying on a more and more natural

978-1-5386-0680-3/17/$31.00 ©2017 IEEE

communicative style, characterized by affective elements, virtual agents have a unique ability to influence humans. A good testbed is provided by the medical domain, where they have proved to be effective promoters of lifestyle changes in human patients [2], [3], or by the education domain, where they can successfully play the role of tutors [4]. The ability of virtual agents to promote beliefs and behaviours in human users is obtained through the creation of an empathic relationship with the user, assuming their goals and emotions. Among the affective elements of the agents' interaction with the users, then, specific attention must be devoted to the so-called moral emotions. Often identified with the Disdain-Rage-Disgust triad [5], but extensible also to the (self-oriented) emotions of shame and pride and to the (other-oriented) emotions of admiration and gratitude, moral emotions are characterized by an intrinsic social nature, and are typically activated based on the values and norms of the society to which an individual belongs. Recognizing and generating moral emotions requires the integration into the agent of the knowledge about the moral aspects of behavior. However, incorporating moral values into a virtual agent is problematic from the ethical point of view, since they can increase the agent's persuasive strategies towards the users.

The paper is structured as follows. Section 2 outlines opportunities for analysis of opinions by NLP techniques. In Section 3, we discuss ethically dangerous relationships between social analysis supported by NPL and easy access to personal user profile data. In Section 4, we discuss the risks and opportunities of the presence of human bias and prejudices in social media corpora. In Section 5, we extend our investigation to the ethical implications of conversations with virtual agents, to end with some conclusions.

2. Love and Hate in the Time of Big Data

By their very nature, social media data fall within the category of big data, not only because of the huge amount of information they are made of, but also because they are representative of society as a whole and across the great variety of facets that manifest in it. Just as a peculiar source of information about society, data extracted from social media are a valuable resource to know the world we live in and the communication dynamics we are daily involved with. Nevertheless, given the prevalence of textual content contained therein, they also represent a major challenge for technologies that deal with human language to extract meaning and knowledge. In fact, although humans are able to understand language as it appears in texts of any kind, even when they are dotted with mistakes, prams and new and creative forms, to build the necessary templates to represent and formalize texts may be extremely challenging, as it is very difficult to build analytical systems working on them. Techniques built within the NLP have so far been compared with texts where compliance with grammatical rules is a prevalent feature, being any other behaviour interpreted as erroneous or marginal. Dealing with texts in social media instead means addressing the challenge of understanding how natural language comprehension passes through our sole morpho-syntactic capabilities and on the contrary requires other types of knowledge that we may not yet know how to formalize. It is not infrequent indeed to notice similarities between spoken language and that used in chat, microblogging and social media.

For what concerns methodology, NLP tools increasingly apply statistical approaches and machine learning techniques, acquiring the most of knowledge directly from data exemplifying phenomena they must deal with. However, this process is usually facilitated by a careful selection of the examples, so that they are representative, and labelling, to make explicit the knowledge they contain.Also in this context comes into play the consideration of ethical aspects, both when selecting data samples, and when designing and applying their annotations. For example, the choice of the data sample will significantly affect the results we will be able to draw by a system trained on them, and the annotation itself must be designed and done in order to avoid biases and to expose the shared knowledge of the speakers' community. However, providing that humans are inclined to express their subjective abilities also in annotation tasks, it should be more appropriate to consider annotated data as representative of a certain type of person rather than of an average speaker, who perhaps only ideally exists.

2.1. Sentiment Analysis and Opinion Mining

In recent years there has been an increasing interest for automatically extracting information about subjective and non-factual content from social media users. Current tools for sentiment analysis, opinion mining and emotion detection refer to different levels of granularity and focusses [6], [7], ranging from the positive or negative polarity of sentiment to the identification of basic emotions [8].

Over time, the generic sentiment analysis task has been outlined in more specific tasks, such as the aspect-based sentiment analysis, focussing on polarities associated with aspects of an entity or stance detection, e.g in the context of political debates, where one tries to identify a specific orientation towards a target entity towards which user opinions tend to polarize [9]. These technologies have recently been applied both in the business sphere and in the analysis of mood in the political sphere or on topics of socio-political interest [10]. In the field of social sciences, instead, it is becoming common to construct, through the analysis of sentiment in social media, indirect measures of subjective well-being which can complement what is obtained through traditional direct survey-based techniques. The final aim is a wise combination of sentiment automatic measurements related on a certain geographic area and traditional socio-economic data on the same populations and territories [11].

A further aspect to be taken into account in the automatic analysis of sentiment is the use of figurative language [12], [13]. Often, sarcastic messages are the ones that spread more virulently. In addition, the presence of ironic expressions can cause the so-called polarity reversal phenomenon, which can undermine the accuracy of sentiment analysis systems.

2.2. Identifying and Monitoring Hate Speech

As a privileged place for expressing any kind of opinions and feelings, social media are also used to convey expressions of hostility and hate speech, as mirrors of social tensions that arise in relation to various events and situations. Also due to the fact that social media support an incredibly wide and rapid spread of messages, the expressions of extreme verbal violence and their proliferation in the network are progressively being considered as indispensable social emergencies to be addressed, by means of coordinated interventions between institutions within individual countries or at the level of broader transnational communities. However, to reach a universally accepted definition of hate speech is still problematic. It has been not easy to define a set of ethical positions acceptable for a large portion of the population, also considering the transnational nature of network communication, and the hate speech notion has been transposed into the laws of the various countries in different ways.

A possible definition of hate speech that sums up the features commonly embedded in this notion, and which can be useful in a computational perspective, includes the presence of three requirements: (i) a clear will and intention to incite hatred, (ii) the presence of a real form of incitement to hatred and violence and (iii) the implementation (or the high risk of implementation) of the violent acts. This means that, in order to recognize the presence of hate speech, it is essential that incitement to hatred is strictly connected with the idea of harm, discrimination or violence itself. Online hate is expressed in different forms depending on the subject against whom it is targeted, and hence can decline into homophobia, racism, xenophobia and hate against migrants. Most of the actions and campaigns carried out against hate speech are aimed to raise awareness among the various components of the population, often based on the monitoring of the phenomenon. The effects of speech hate are particularly known in the weakest sections of the population, such as young people. Sentiment analysis technologies are particularly suitable for detecting and monitoring speech hate [14]. They make it possible to collect data of varying nature, present in texts extracted from social media or in related metadata. These data can be aggregated to represent the hate phenomenon for example in the form of maps [15].

3. Dangerous Liaisons: What is Ethical to Infer from the Social Data that we Analyze?

Particular attention must be paid to the confidentiality of certain data and to protect the individual by applying appropriate forms of anonymisation. Nevertheless, intervening *ex-post*, when analyzing and aggregating data already produced by users, does not seem sufficient, and practices devoted to improve the awareness *ex-ante* of users should be applied. When social media data are analyzed, we are indeed in a dimension in which private and public spheres mingle and this leads to a further level of ethical considerations to be taken into account in order to properly address.By recognizing user's profile, personality traits, and other features, not only activities devoted to guarantee citizen security but also unwanted or undesirable investigations can be done finalized at committing unlawful or unethical acts against them. For example, there are techniques such as *author profiling* that allow, though with a certain margin of error, to discover the demographic characteristics of the author of a social media media [16] or even his personality traits [17]. The same technologies can be applied for harming morally or economically disadvantaged weak subjects, like manipulating consumer behaviors.

Also for what concerns the political domain, we must be very cautious about the possibility of using social media data to predict the results of elections [18]. Not all information in social media can be considered trusted. The presence of propaganda and misinformation, viral hoaxes must be taken into account. Representativeness of social data from the point of view of the demographic categories is another important issue which can heavily affect the analyses: not all ages, genres and social groups are equally represented. Lastly, one should consider self-selection bias, which sees the most active users most politically present in the debate. But one the most worrying ethical implications on this side concerns the possibility to use the data on the personality of users to manipulate political sentiment in favor of a political entity. Some recent journalistic inquiries have shed light on the important role in the latest US presidential campaign and the English Brexit referendum of new companies that deal with data analysis and strategic communication in electoral processes. The underlying idea behind the approach of such companies is that it is possible, by determining the users' personality based on their textual traces on social media and combining this information with sentiment analysis of their tastes and traditional statistical data, guessing the voting intentions of users, and better calibrate, personalizing it, the campaign of the still undecided voters.

Similar issues are involved in hate speech analyses. Since the boundary between censorship of speech hate and freedom of expression restriction is very subtle [19], it is possible to manipulate the political purpose of monitoring hate speech. Further ethical implications arise from the automatic analysis of non-textual content in social media, which increasingly support multimodal communication with texts and images together. The applications of modern and sophisticated machine learning techniques based on the neural paradigm to large datasets, for extracting information from images, are leading to make considerable progresses. Automatic systems can analyze the physical characteristics of a person based on their image. As pointed out in [20], from here to the possible establishment of a technologically supported "new physiognomy", where the goal is extracting from the image analysis of a face information about the character, the personality and the possible tendency to implement criminal behaviour, the pace is short. Both researchers and the technology companies must be well-equipped with awareness and sense of responsibility, provided that machine learning models tend to incorporate the bias present in human behavior.

4. Human Bias, Prejudices and Stereotypes in Social Media Corpora

The ‘ecological assumption’, which characterizes the corpus-based approaches to linguistic analysis from the beginning, naturally leads to the idea that the meanings extracted from texts belonging to a particular domain or environment are intrinsically imbued with cultural stereotypes, conventions, and values broadly shared in the community of users expressing themselves in the corpora [21]. Corpora of texts from social web users contain recognizable footprints of the historical prejudices of the users community, which in this context, however, express themselves using a peculiarly direct and unfiltered style, in a communicative context that inhibits our very human capacity to be empathic typical of interpersonal contexts [22]. Prejudice is an unfavorable attitude towards people or groups of people. According to recent theories of social psychology [23], which attribute an important role to the affective component of prejudice and pay particular attention to the analysis of its hidden or implicit forms [24], prejudice consists of several elements: stereotypes; expression of negative emotions such as antipathy, hostility, hatred; implementation of hostile and discriminatory behaviors towards members of a group or social category, motivated only by the fact that they are belonging to that group or category. As well-emphasized in a recent study [1], the machine learning algorithms that NLP systems rely on for extracting semantic information from corpora, when trained on spontaneous texts produced by a community which, even implicitly, expresses a prejudicial attitude towards people or groups of people, learn a semantics that is imbued with the stereotypes and the human bias underlining that community and society. This has important implications from an ethical point of view. First, it warns us of the danger of developing applications that, on the one hand, acritically absorb the worst cultural stereotypes expressed by our societies and, on the other hand rely on this knowledge, automatically learned from language, in the subsequent process of planning their interactions with the user and with the real world. This is an element to keep in mind, especially in the context of the increasingly widespread use of AI tools, that exploit these technologies to develop applications that are delegated to take decisions affecting our society of people. A thread of ethical reflections of the scientific community is recently flourishing on this issue, both with regard to NLP [25] and other areas of artificial intelligence where the focus is, more in general, about the theme of so-called “fairness” in machine learning. Secondly, the awareness about the presence of these bias and the ability to automatically study and identify them [1], can offer a precious tool for psychology and social sciences to study different types of human bias on appropriately selected linguistic corpora. Following this line, there is the possibility to thoroughly analyze and monitor phenomena of hatred and hostility towards certain groups of peoples, and to shed light on stereotypes and prejudices, also producing implicit measures of prejudice similar to those traditionally obtained through the well-known Implicit Association Test [24], by exploiting computational models of distributional semantics based on word embedding [26], as proposed in [1] through the Word Embedding Association Test. Such analyses may be an important premise in order to outline actions to contrast prejudice and hate speech. For example, analyzing online hate and, together, the implicit perceptions of widespread hostility towards migrants, it would be possible to acquire actionable knowledge for society and politics, to design new actions to foster integration, and to find new ways to address one of the challenges of the century [27].

5. The Relationship between Purpose, Moral Values, and Persuasion in Virtual Agents

Virtual agents, especially if equipped with a graphical interface that enables the use of multiple channels (voice, gesture, gaze, posture, etc.) in the communication with the user, have a high potential in the development of assistive applications, thanks to their ability to establish a natural interaction with the user [28]. The integration into artificial agents of the ability to feel – and hence show – human emotional states and to understand the emotions of the user, being empathic to her/him, increases the agents’ ability to become credible partners of humans. This ability is particularly precious in areas characterized by the need to establish a relationship of trust and attachment with the user, such as learning or medical care [29]. The nature of this type of interaction, coupled with the continuity of the relationship over time, can be ethically problematic, especially if it involves categories of people with fragility, such as children elderly or sick people. As pointed out by the research field of captology, namely the study of persuasion through computer technology [30], there is a fine line between well meant persuasion and safeguard of the user’s free will. The emotions of the user become the input to the deliberation of the virtual agent, who can leverage them to achieve its built-in purposes: the agent can help the user overcome the negative emotions that obtrude the achievement of her/his own goals, or use her/his positive emotions to help the her/him to achieve these goals. For example, in the promotion of healthy lifestyles (such as compliance with a particular dietary plan), the user’s emotion of shame for actions that are not in line with the diet (eating foods that are not in the plan) may be exploited by the virtual agent to orient the user towards virtuous actions in the future; similarly, the user’s emotion of happiness for complying with the diet may lead the agent to devise strategies that enhance the user’s motivation (e.g., by congratulating the user for the achieving the target). If the boundary between the benefits for the user and abuse of persuasion can be rather easily detected in the example above, it is not easy to find general criteria for all domains, given the variety of applications that potentially involve virtual agents.

A closer look to the architecture of emotional virtual agents, however, suggests a way to address the ethical aspects in their persuasive behaviour. The emotional input provided by the user, conveyed to the agent through the means available in the specific type of interaction (e.g., voice, text, posture, gestures, and facial expressions), is processed by the agent, which maps them onto emotional states in the user model. In particular, based on a long tradition of research dating back to Ekman's seminal work on basic emotions [31], facial expression has nowadays become of primary importance, with off-the-shelf technologies for the recognition of emotions from facial expression available to any developers. These techniques leverage datasets that encompass a large number of facial expressions annotated with the emotion categories they represent, together with anatomic-morphological information [32]. To overcome of the limitations of the so-called 'primary emotions', scholars have resorted to models developed by cognitive psychology to grasp the complex interaction between emotions and intentional behaviour. While the notion of emotional appraisal refers to the activation of emotional states in response to the dynamic relationship between the individual and the environment [33], the notion of emotional coping refers to the response to emotions aimed at reducing their impact on the individual [34]. Over the last decade, appraisal and coping processes have been integrated into virtual agents' architectures [35], yielding sophisticated software entities that mimic human emotions in a believable way: in those architectures, the emotion appraisal module generates the agent's emotional states based on the relationship between its goals and the environment (that facilitates or hinders their achievement); coping processes mediate between the emotions generated by the appraisal process and the agent's deliberation, making it react to its own emotions. With the ability to process the emotions of the users and understand (and possibly assume) their goals, virtual agents can behave empathetically and affect the behaviour and emotions of the users.

The modularity of these processes in agent architecture, explicitly moulded on human cognition, can be the key to a solution to ethical issues. The agent's emotional models, in fact, rely in most cases on some explicit description of the agent's goals and values, and on their relation with the users' inferred goals: goals enable the agent to devise plans to achieve its goals, and their achievement is the input to the emotional appraisal; some encoding of values is needed for the agent to assess the moral consequences of its own and others' behaviour [36], and to generate moral emotions. In addition, the agent equipped with the mechanisms described above can use the same appraisal and coping processes to infer the user's emotional states. The need to incorporate into the agent an explicit description of its goals and values, motivated by the implementation of appraisal and coping processes, opens the way to identifying these goals and values with no ambiguity, allowing the designers and developers to control the behaviour of the agent in an ethical sense. So, the controversies raised by the relationship between goals, moral values, and persuasion in virtual agents [37], can be at least partially addressed by the separation of the moral competence, goals and deliberative mechanisms that characterizes most agent models and architectures. Agent architectures, in fact, typically rely on declarative constructs such as goals, intentions, and norms which, inspired by the mentalistic model of behaviour provided by Bratman [38], enable the programming of agents for practical applications. The ability to identify (and configure) the elements that drive the agent's behaviour - the goals the agent assumes, the principles guiding its moral component - is the key to the development of artificial agents whose purposes are transparent and explicit and can be fully scrutinized from outside. The presence of the declarative components mentioned above, expressed in a shared semantics derived from [38], make it possible to verify and discuss the moral implications of virtual agents, and to devise ways to resolve the complex interaction of goals and values, for example by resorting to explicit prioritization as suggested by van Fraassen [39]. In this sense, the adoption of sub-symbolic approaches to the generation of the communicative behavior of agents [40] can put at stake the verification of the ethically sensitive components of the agent.

6. Conclusion

The reflection path presented in this paper has as main long-term goal the development and growth of a deeper awareness about the ethical aspects related to the application of language and communication technologies to the analysis of large amounts of data generated by social media users. We observe and try to understand in this context dynamics of virtuous interaction, but also hostility and hate, that we want to monitor in order to extract useful knowledge to act or counteract them. The great opportunities we have outlined here must be considered taking into account the ethical implications concerning *dangerous liaisons* between the information we extract by analyzing users' spontaneous expressions and the digital traces that relate to the privacy spheres of their personal lives, but also machine learning models used for automatic analysis of these data which tend to incorporate bias present in human behavior.Although not intentional, this kind of propagation of human prejudice through automatic algorithms can present some risks, if the phenomenon is not recognized.

When we expand our look to include communication with virtual agents, it can be seen how the standard architecture of such systems naturally allows to identify the critical elements for the ethical dimension. However, if it is for the scientific community to tell where such elements are, it is up to national and supranational authorities to use such knowledge to develop guidelines that allow an ethical exploitation of the great potential of such technologies.

Acknowledgments

C. Bosco and V. Patti were partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01) and by Fondazione CRT (*Hate Speech and Social Media*, 2016.0688).

References

[1] A. Caliskan, J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human biases," *Science*, vol. 356 (6334), pp. 183–186, 2017.

[2] T. Bickmore, R. Asadi, A. Ehyaei, H. Fell, L. Henault, S. Intille, and C. Shanahan, "Context-awareness in a persistent hospital companion agent," in *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, 2015, pp. 332–342.

[3] C. Lisetti, R. Amini, U. Yasavur, and N. Rishe, "I can help you change! an empathic virtual agent delivers behavior change health interventions," *ACM Transactions on Management Information Systems (TMIS)*, vol. 4(4), 2013.

[4] G. Castellano, A. Paiva, A. Kappas, R. Aylett, H. Hastie, W. Barendregt, and S. Bull, "Towards empathic virtual and robotic tutors," in *Proc. of the International Conference on Artificial Intelligence in Education*. Springer, 2013, pp. 733–736.

[5] P. Rozin, L. Lowery, S. Imada, and J. Haidt, "The cad triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity)." *J. of personality and social psychology*, vol. 76, no. 4, p. 574, 1999.

[6] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2(1-2), pp. 1–135, 2008.

[7] C. Strapparava and R. Mihalcea, "Affect detection in texts," in *The Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds. Oxford University Press, 2015.

[8] M. Nissim and V. Patti, "Semantic aspects in sentiment analysis," in *Sentiment Analysis in Social Networks*, F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, Eds. Morgan Kaufmann, 2017, pp. 31–48.

[9] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Trans. Internet Technol.*, vol. 17, no. 3, 2017.

[10] C. Bosco and V. Patti, "Social media analysis for monitoring political sentiment," in *Encyclopedia of Social Network Analysis and Mining*, R. Alhajj and J. Rokne, Eds. Springer, 2017.

[11] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth, "The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place," *PLoS ONE*, vol. 8, no. 5, 2013.

[12] J. Karoui, B. Farah, V. Moriceau, V. Patti, C. Bosco, and N. Aussenac-Gilles, "Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study," in *Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: ACL, 2017, pp. 262–272.

[13] E. Sulis, D. I. Hernández Farías, P. Rosso, V. Patti, and G. Ruffo, "Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not," *Knowl.-Based Syst.*, vol. 108, pp. 132–143, 2016.

[14] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. of the 5th International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: ACL, 2017, pp. 1–10.

[15] C. Musto, G. Semeraro, M. de Gemmis, and P. Lops, "Modeling community behavior through semantic analysis of social data: The italian hate map experience," in *Proc. of the 2016 Conference on User Modeling Adaptation and Personalization*, 2016, pp. 307–308.

[16] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd author profiling task at PAN 2015," in *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, ser. CEUR Workshop Proceedings, vol. 1391, 2015.

[17] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[18] D. Gayo-Avello, "I wanted to predict elections with Twitter and all i got was this lousy paper," *CoRR*, vol. abs/1204.6441, 2012. [Online]. Available: http://arxiv.org/abs/1204.6441

[19] C. Yong, "Does freedom of speech include hate speech?" *Res Publica*, vol. 17, no. 4, p. 385, Jul 2011.

[20] B. Agëra y Arcas, M. Mitchell, and A. Todorov, "Physiognomys new clothes," https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a, 2017.

[21] M. Stubbs, *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture*, ser. Language in Society. Wiley, 1996.

[22] S. Turkle, *Reclaiming Conversation: The Power of Talk in a Digital Age*. Penguin Publishing Group, 2015.

[23] R. Brown, *Prejudice: Its Social Psychology*. Wiley, 2011.

[24] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz, "Measuring individual differences in implicit cognition: The implicit association test," *Journal of Personality and Social Psychology*, vol. 74, pp. 1464–1480, 1998.

[25] D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, and H. Wallach, Eds., *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: ACL, 2017. [Online]. Available: http://www.aclweb.org/anthology/W17-16

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[27] C. Bosco, V. Patti, M. Bogetti, M. Conoscenti, G. Ruffo, R. Schifanella, and M. Stranisci, "Tools and resources for detecting hate and prejudice against immigrants in social media," in *Proc. of 1st Symposium on Social Interactions in Complex Intelligent Systems (SICIS), AISB Convention 2017, AI and Society*, Bath, UK, 2017.

[28] M. Ochs, C. Pelachaud, and D. Sadek, "An empathic virtual dialog agent to improve human-machine interaction," in *Proc. of the 7th international joint conference on Autonomous Agents and Multiagent Systems*. IFAAMAS, 2008, pp. 89–96.

[29] M. L. Hoffman, *Empathy and moral development: Implications for caring and justice*. Cambridge University Press, 2001.

[30] B. Fogg, "Captology: the study of computers as persuasive technologies," in *CHI'97 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1997, pp. 129–129.

[31] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[32] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[33] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, "Appraisal theories of emotion: State of the art and future development," *Emotion Review*, vol. 5, no. 2, pp. 119–124, 2013.

[34] S. Folkman and R. S. Lazarus, "Coping as a mediator of emotion." *J. of personality and social psychology*, vol. 54, no. 3, p. 466, 1988.

[35] J. Gratch and S. Marsella, "A domain-independent framework for modeling emotion," *Cognitive Systems Research*, vol. 5, no. 4, pp. 269–306, 2004.

[36] C. Battaglino, R. Damiano, and L. Lesmo, "Emotional range in value-sensitive deliberation," in *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. IFAAMAS, 2013, pp. 769–776.

[37] S. Stark, "A change of heart: Moral emotions, transformation, and moral virtue," *J. of Moral Philosophy*, vol. 1, no. 1, pp. 31–50, 2004.

[38] M. Bratman, "Intention, plans, and practical reason," 1987.

[39] B. C. Van Fraassen, "Values and the heart's command," *The Journal of Philosophy*, vol. 70, no. 1, pp. 5–19, 1973.

[40] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep reinforcement learning for dialogue generation," *arXiv preprint arXiv:1606.01541*, 2016.