

Automatic Detection of Hope Speech

Daniel García-Baena

Computer Science Department, SINAI research group, CEATIC, Universidad de Jaén, Spain

Abstract

Hope speech is a type of discourse that has the power to help, inspire people for good and even relax hostile environments. The automatic detection of hope speech is an open challenge in Natural Language Processing that has been generally eclipsed by hate speech detection. Rather than simply deleting hate speech from the Internet, restricting freedom of speech, according to the outstanding importance that psychology gives to hope and the success that some social experiments had when they highlighted hope speech over the rest of the texts, we find specially necessary to study in depth the automatic identification of hope speech. In this work, we describe a thesis project that focuses on the development of new datasets and systems that allow the automatic detection, by means of different classical machine learning techniques and new deep learning architectures, of hope speech, mainly in Spanish.

Keywords

Hope speech, natural language processing, language that relaxes hostile environments, language that promotes equality, diversity and inclusion

1. Justification of the research

Hope speech is the type of speech that is able to relax a hostile environment [1] and that helps, gives suggestions and inspires for good to a number of people when they are in times of illness, stress, loneliness or depression [2]. Detect it automatically, so that positive comments can be more widely disseminated, can have a very significant effect when it comes to combating sexual or racial discrimination or when we seek to foster less bellicose environments [1].

As stated in the work of Chakravarthi [2], hope speech is defined as the language that is related to fostering individuals' potential, supporting them and reaffirming their self-confidence, as well as, again, making motivational and inspirational suggestions in difficult times of illness, loneliness, stress or depression [3].

However, Palakodety et al. [1] differ from the above definition and establish as hope speech simply that which has the capacity to relax situations of tension and violence. Even, Chakravarthi [2] also introduces a possible variation of what is meant by hope speech, now taking into account the ability of language to promote equality, diversity and inclusion (EDI) of women belonging to the fields of science, technology, engineering and management (STEM), lesbian, gay, bisexual, transgender, intersex and queer individuals (LGBTIQ); and racial minorities and individuals with disabilities.


Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA, 28 September 2023, Jaén, Spain.

✉ daniel.gbaena@gmail.com (D. García-Baena)

🆔 0000-0002-3334-8447 (D. García-Baena)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In this thesis it is pretended to elaborate resources in order to automatically classify hope speech. Therefore, it will be created a new Spanish written dataset for hope speech identification and it will be developed too some systems for detecting hope speech.

2. Previous works

As this is a recent task to be tackled automatically from Natural Language Processing (NLP), only a few corpora are available. Until now, the work that has been done in relation to hope speech identification has been focused in developing new datasets for English, Malayalam and Tamil; and automatic detection systems based on classic machine learning strategies and modern deep learning architectures. They will be discussed below.

2.1. HopeEDI

The HopeEDI dataset [2] contains comments in English, Malayalam and Tamil. It consists of data obtained from comments posted on YouTube videos that were collected from November 2019 to June 2020. The corpus can be downloaded free of charge at Hugging Face: https://huggingface.co/datasets/hope_edi.

The subject matter of the comments written in English is EDI (Equality, Diversity and Inclusion). In this case, the comments come from videos posted by Indian and Sri Lankan users. It is important to note that since India is a multilingual country, many of the comments may be written in several languages at the same time (code-mixing).

For the HopeEDI corpus, its author applied different machine learning algorithms on a TF-IDF (Term Frequency-Inverse Document Frequency) representation of the tokens. Specifically, the corpus was evaluated with the following: Bayesian multinomial classifier (multinomial Naïve Bayes or MNB) with a value of *alpha* equal to 0.7, k-nearest neighbors method, support vector machine (SVM), decision tree (DT), and with Logistic Regression (LR). In any case, for all commented techniques, results scored an F1 value no better than 0.56 and, consequently, they were quite disappointing.

2.2. India-Pakistan

This dataset contains data from English comments posted on videos from YouTube [1]. The researchers chose this site as the source of the data because it is the most widely used video broadcasting platform in India and Pakistan today. Unfortunately, this dataset is not publicly available.

For their compilation, a series of queries were prepared and then extended with searches related to the Kashmir conflict by consulting trends from India and Pakistan that took place between February 14, 2019 and March 13, 2019. Finally, such queries were used to search for related videos on YouTube and subsequently obtain their comments using the public API of that social network.

The comments are all written in English and come from mainly Indian and Pakistani users. There are also comments submitted by immigrants from India and Pakistan, whose were in Bangladesh, Nepal, United States, United Kingdom, Afghanistan, China, Canada and Russia. In

this case, the origin of the users was taken into account with the intention of maintaining an equal representation of citizens belonging to both sides of the conflict.

This time, the authors used a logistic regression with L2 regularization classifier (Ridge Regression). The experiment was run a total of one hundred times on one hundred random sections of the dataset and achieved an F1 value of 0.79.

2.3. KanHope

KanHope dataset [4] contains comments in code-mixed Kannada-English. All data was collected with the app YouTube Comment Scraper between February 2020 and August 2020. The dataset is publicly available on Hugging Face: https://huggingface.co/datasets/kan_hope.

KanHope gathers comments from several videos on distinctive topics such as movie trailers, India-China border dispute, people’s opinion about the ban on several mobile apps in India, Mahabharata and other social issues that involved oppression, marginalization and mental health. KanHope dataset authors emphasize on the inclusion of people of marginalized communities, such as LGBT, racial and gender minorities. All comments were from users based in India and, being it a multilingual country, researchers were motivated to extract the comments to work on code-mixed texts.

The corpus authors applied from primitive machine learning to complex deep learning approaches. The model DC-BERT4HOPE (roberta-mlm) obtained the best results for F1-scores with 0.752, followed by DC-BERT4HOPE (bert-mlm): 0.735, mBERT: 0.726, DC-BERT4HOPE (roberta-xlm): 0.720, and random forest with 0.706.

2.4. SpanishHopeEDI

Finally, we have generated a quality dataset SpanishHopeEDI [5], a new Spanish Twitter corpus on LGBT community, and we have conducted some experiments that can serve as a baseline for further research. The dataset consists of 1,650 LGBT-related tweets annotated as HS (Hope Speech) or NHS (Non Hope Speech). A tweet is considered as HS if the text:

1. Explicitly supports the social integration of minorities.
2. Is a positive inspiration for the LGTBI community.
3. Explicitly encourages LGTBI people who might find themselves in a situation or unconditionally promotes tolerance.

On the contrary, a tweet is marked as NHS if the text:

1. Expresses negative sentiment towards the LGTBI community
2. Explicitly seeks violence or uses gender-based insults.

The dataset was created from LGBT-related tweets. All of those tweets were written in Spanish and were collected using the Twitter API. As seed for the search we used a lexicon of LGBT-related terms, such as #OrgulloLGTBI and #LGTB. In addition, it should be mentioned that our SpanishHopeEDI dataset was included in the second workshop on Language Technology for Equality, Diversity and Inclusion that was held as a part of the ACL 2022 [6].

3. Description, hypotheses and objectives

EDI is an important issue in many different areas. Language is a fundamental tool for communication and it must be inclusive and treat everyone equally. However, sometimes on social media this is not the case, as more offensive messages are posted towards people because of their race, color, ethnicity, gender, sexual orientation, nationality or religion. As Chakravarthi [2] stated, the importance of the social media on the lives of vulnerable groups, such as for people belonging to the LGBT community, racial minorities or individuals with disabilities; plays an essential role in shaping their personalities and how they perceive society [7, 8, 9]. Therefore, it is found important to focus on researching on the inclusion of this people and to use promoting positive content on social media, in pursuit of EDI.

The importance of hope has already been carefully studied by psychologists and, consequently, we can affirm that hope plays a crucial role in the well-being, recovery and restoration of humans [2]. Greater hope is consistently related to a better academic, athletic, physical health, psychological adjustment and psychotherapy outcomes. In general, Hope Theory is comparable to theories of Learned Optimism, Optimism, Self-Efficacy and Self-Esteem [10].

Individuals with high doses of hope do not react in the same way to barriers as those with low amounts of hope, but instead view barriers as challenges to overcome and use their pathway thoughts to plan an alternative route to their goals [11, 12]. In addition, high levels of hope has been found to be correlated with a number of beneficial elements, such as academic performance [13] and lower levels of depression [14]. In contrast, low hope proportions are associated with negative outcomes, such as reduced well-being [15].

Therefore, it is relevant to analyze the state of the art of automated hope speech detection technologies from the perspective of NLP. In this sense, automated detection of hope speech can be especially useful in promoting the dissemination of hopeful messages to those in difficult times and can be used to promote positive messages to support EDI. Previous studies have shown that a snowball effect occurs in social media and abusive comments lead to more abusive comments and positive comments inspire people to leave more positive comments [16, 17]. In order to study this, Facebook conducted an experiment by modifying its *Newsfeed* algorithm to show more positive or negative posts to certain users [18]. Their results showed that people tend to write positive posts when they see happy posts in their newsfeeds and vice versa. All this suggest the importance of reinforce positivity on social media, focusing then on promoting hope speech.

Hence, it was considered important to pursue the following objectives:

1. To theoretically study the concept of hope speech, as well as its treatment from an NLP point of view.
2. Analyzing the already existing hope speech detection solutions and discussing the problems derived from them.
3. To make a review of all available resources, providing experiences and an accessible introduction to those researchers who may be interested in tackle this problem.
4. Make a new dataset focused on the LGBT community for Spanish hope speech detection.
5. Create baseline experiments using machine learning and deep learning algorithms, including, of course, cutting edge technologies as transformers models.

6. Develop an extensive error analysis in order to be able to determine future directions of this study.

4. Methodology

The methodology that is proposed in order to achieve the objectives of this thesis is detailed below:

1. Firstly, it is necessary to carefully review the state of the art of hope speech classification. Therefore, it will be important to evaluate both already existing corpus and classification systems.
2. Secondly, we will part from some of the currently available resources, in relation to hope speech detection, and we will develop new ones with the intention of make it possible to detect hope speech sentences from texts written in Spanish.
3. Therefore, it will be created a new corpus, containing several texts written only in Spanish, that we will focus in EDI.
4. Then, we will create different systems that will use the last dataset for making possible to automatically identify hope speech texts.
5. And, finally, we will experiment with and evaluate our new resources so as to improve them, always sharing our work with the scientific community, publishing all the results and organizing shared tasks.

5. Research questions

The main research questions that we pretend to respond with this work are all of them listed afterwards:

- How similar it is to detect hope and hate speech?
- It is possible to elaborate unambiguous hope speech tagging notes?
- Are tagging notes for hope speech corpus dependent of the language in which the texts from the dataset were written?
- It is interesting, or useful, to create hope speech datasets for making possible to automatically detect it?
- What can we learn from the already existing datasets for hope speech detection in languages different than Spanish?
- For new classification systems, how could we improve them?
- In relation to hope speech detection, is it viable to identify the main causes of possible classification errors?
- What algorithms are the best for automatic detection of hope speech?

Acknowledgments

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government, Project FedDAP (PID2020-116118GA-I00) supported by MICINN/AEI/10.13039/501100011033, WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government and by a grant from Fondo Social Europeo and the Administration of the Junta de Andalucía (DOC_01073).

References

- [1] S. Palakodety, A. R. KhudaBukhsh, J. G. Carbonell, Hope speech detection: A computational analysis of the voice of peace, arXiv preprint arXiv:1909.12940 (2019).
- [2] B. R. Chakravarthi, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: <https://aclanthology.org/2020.peoples-1.5>.
- [3] C. R. Snyder, S. J. Lopez, H. S. Shorey, K. L. Rand, D. B. Feldman, Hope theory, measurements, and applications to school psychology., *School psychology quarterly* 18 (2003) 122.
- [4] A. Hande, R. Priyadharshini, A. Sampath, K. P. Thamburaj, P. Chandran, B. R. Chakravarthi, Hope speech detection in under-resourced kannada language, 2021. arXiv:2108.04616.
- [5] D. García-Baena, M. García-Cumbreras, S. M. Zafra, J. García-Díaz, R. Valencia-García, Hope speech detection in spanish, *Language Resources and Evaluation* (2023) 1–28. doi:10.1007/s10579-023-09638-3.
- [6] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Chinnaudayar Navaneethakrishnan, J. P. McCrae, M. A. García-Cumbreras, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumar Kumaresan, R. Ponnusamy, D. García-Baena, J. A. García-Díaz, Overview of the shared task on hope speech detection for equality, diversity, and inclusion, *Association for Computational Linguistics* (2022) 378–388. URL: <https://aclanthology.org/2022.ltedi-1.58>. doi:10.18653/v1/2022.ltedi-1.58.
- [7] V. Kitzie, I pretended to be a boy on the internet: Navigating affordances and constraints of social networking sites and search engines for lgbtq+ identity work, *First Monday* (2018).
- [8] P. Burnap, G. Colombo, R. Amery, A. Hodorog, J. Scourfield, Multi-class machine classification of suicide-related communication on twitter, *Online social networks and media* 2 (2017) 32–44.
- [9] D. N. Milne, G. Pink, B. Hachey, R. A. Calvo, Clpsych 2016 shared task: Triaging content in online peer-support forums, in: Proceedings of the third workshop on computational linguistics and clinical psychology, 2016, pp. 118–127.
- [10] C. R. Snyder, Hope theory: Rainbows in the mind., *Psychological Inquiry* 13 (2002) 249–275.

- [11] C. R. Snyder, *The psychology of hope: You can get there from here*, Simon and Schuster, 1994.
- [12] C. R. Snyder, Hypothesis: There is hope, in: *Handbook of hope*, Elsevier, 2000, pp. 3–21.
- [13] C. R. Snyder, H. S. Shorey, J. Cheavens, K. M. Pulvers, V. H. Adams III, C. Wiklund, Hope and academic success in college., *Journal of educational psychology* 94 (2002) 820.
- [14] C. R. Snyder, B. Hoza, W. E. Pelham, M. Rapoff, L. Ware, M. Danovsky, L. Highberger, H. Ribinstein, K. J. Stahl, The development and validation of the children's hope scale, *Journal of pediatric psychology* 22 (1997) 399–421.
- [15] E. Diener, Subjective well-being, *The science of well-being* (2009) 11–58.
- [16] A. Sundar, A. Ramakrishnan, A. Balaji, T. Durairaj, Hope speech detection for dravidian languages using cross-lingual embeddings with stacked encoder architecture, *SN Computer Science* 3 (2022) 1–15.
- [17] L. Muchnik, S. Aral, S. J. Taylor, Social influence bias: A randomized experiment, *Science* 341 (2013) 647–651.
- [18] A. D. Kramer, J. E. Guillory, J. T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, *Proceedings of the National Academy of Sciences* 111 (2014) 8788–8790.