

Exploring violence against women in social media through artificial intelligence

Carlos Perrián-Pascual
Universitat Politècnica de València

Abstract

Violence against women remains one of the most pressing social problems in many societies today. Indeed, it is among the most widespread forms of human rights violations, a reality which cuts across social, cultural, religious, economic and geographical boundaries. In this context, the purpose of our study is to determine whether a pre-trained small language model can be leveraged to automatically detect tweets that reflect social concern about violence against women. To this end, we conducted an experiment using two datasets of Spanish and English tweets. Only half of these tweets actually described violence against women as an ongoing social issue, even though all of them contained words typically associated with this problem type. The experiment revealed that a method that we called STECS, which is based on self-organising maps trained on embeddings from synthetic corpora generated by a language model, delivered better classification results than in-context learning based on various prompt-engineering techniques (e.g. zero-shot and chain-of-thought). In particular, our method obtained an average of 0.835 in F1. This chapter sheds new light on unsupervised fine-grained text classification, where the STECS method can be employed to develop real-world applications aimed at detecting community problems (e.g. violence against women) in social media.

Keywords: violence against women, social problem, social media, classification, language model, prompt engineering, neural network

1- Introduction

Violence against women (VAW) is recognised as being one of the most pervasive and severe violations of human rights, which affects millions of people worldwide. As stated by *The Declaration on the Elimination of Violence Against Women*, VAW is defined as “any act of gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life” (United Nations, 1993). *The Declaration* also includes that VAW should be understood to encompass the following:

- a. Physical, sexual and psychological violence occurring in the family, including battering, sexual abuse of female children in the household, dowry-related violence, marital rape, female genital mutilation and other traditional practices harmful to women, non-spousal violence and violence related to exploitation;
- b. Physical, sexual and psychological violence occurring within the general community, including rape, sexual abuse, sexual harassment and intimidation at work, in educational institutions and elsewhere, trafficking in women and forced prostitution;
- c. Physical, sexual and psychological violence perpetrated or condoned by the State, wherever it occurs.

From the previous definition, it follows that “gender-based violence” is a broader term, which refers to “any type of harm that is perpetrated against a person or group of people because of their factual or perceived sex, gender, sexual orientation and/or gender identity” (Council of Europe, 2021). In other words, this term generally refers to harmful behaviours that stem from unequal power relations and are directed at individuals because of socially constructed gender roles (Mooney et al., 2021). Within sociological research, scholars often draw a distinction between gender-based violence and VAW. Whereas the former broadly refers to violence that arises from unequal power relations based on gender roles, the latter specifically focuses on violence where women

are the victims and men are typically the perpetrators (Shepherd, 2008). However, while this distinction is conceptually useful, she also recognises that both terms are often used interchangeably in public discourse, since gender-based violence disproportionately affects women and girls.

In this context, exploring posts published on social-media platforms such as microblogs (e.g. X, formerly known as Twitter), social networks (e.g. Facebook) and forums, where citizens can reflect their concerns about VAW, can contribute to reproducing situation awareness, which refers to knowing what is going on around you in a decision-making process (Endsley, 2000). As social-media posts can be valuable indicators of how VAW is perceived within different communities, analysing these perceptions can help researchers gain insights into this social problem, while policymakers can use the findings to implement corrective measures accordingly. Considering social-media users as witnesses of a given society is in line with social sensing, which involves the analysis of digital communications to detect real-world events or situations in which users act as sensors (Sakaki et al., 2013; Musto et al., 2015; Arthur et al., 2018).

This chapter aims to perform fine-grained text classification in community-problem detection by leveraging a generative AI model.¹ To this end, we conducted an experiment to classify Spanish and English tweets with respect to the social problem of VAW by means of an open-source small language model (SLM), since it can be readily integrated into downstream applications as a cost-effective solution for production en-

1. In the framework of Smart Society, Perrián-Pascual (2023) developed the DIAPASON ontology, which models a large variety of community-problem types from the perspective of citizens. In the hierarchical organisation of this ontology, the upper level is structured into two problem realms: the social realm, which subsumes the LIVING, ECONOMY and GOVERNANCE problem dimensions, and the physical realm, which subsumes the MOBILITY, INFRASTRUCTURE and ENVIRONMENT problem dimensions. In this regard, VAW is ontologically categorised as a community problem in the social realm.

vironments. The main contribution of this study lies in exploring two methods in which the SLM is involved: (a) in-context learning with prompt-engineering techniques (i.e. zero-shot and chain-of-thought) and (b) unsupervised neural networks (e.g. self-organising maps) trained with synthetically generated corpora. Our study concludes that the latter serves as an effective method for unsupervised single-label binary classification in detecting this social problem. We say that this type of classification is “unsupervised” because it does not require an annotated dataset to train the model, “single-label” because each tweet in the test datasets is assigned only one label in the learning process, and “binary” because it involves predicting one of two possible classes (i.e. whether or not the tweet describes the problem of VAW).

The remainder of this chapter is organised as follows. Section 2 and Section 3 give an overview of the study of social problems and pre-trained language models, respectively. Section 4 provides an account of the proposed research methodology. Section 5 deals with the experiment of this study. Finally, Section 6 presents some conclusions.

2- A constructionist view of social problems

There are two main ways to study social problems. One approach, often called “objectivist”, identifies social problems with conditions that are harmful to society (Davis, 1937). The other approach, known as “subjectivist”, identifies social problems with conditions that people perceive as problematic (Spector and Kitsuse, 1977; Best, 1989; Holstein and Miller, 2003). While the objectivist approach still dominates most sociology textbooks (DeFronzo and Gill, 2020), the subjectivist perspective has been gaining interest in recent years. One reason for this is that objectivist definitions often leave the concept of “harm” poorly specified, making it difficult to apply meaningfully (Best, 2017). Furthermore, not all harmful conditions can be recognised as social problems, particularly if they do not attract enough public attention. As DeFronzo and Gill (2020) noted, whether a condition is treated as a social problem

depends not only on its existence but also on the level of public concern it generates.

Given these considerations, our study adopts the subjectivist approach to social-problem detection, where a certain condition becomes a social problem when people react negatively to it. Therefore, we have to redirect the focus from the conditions themselves to the claims people make about them. In other words, the emphasis is on the social processes through which problems are recognised, a concept sociologists refer to as social construction (Berger and Luckmann, 1966). As a result of this constructionist view, Best (2017) introduced claimsmaking (i.e. the act of bringing the problem to public attention by making a claim) as the first stage for studying social problems. Claimsmaking “determines both which phenomena will be designated as social problems, and which characteristics are attributed to those problems” (Letukas, 2014, 49). In this regard, three elements are central in claimsmaking: claims, claimsmakers, and troubling conditions. Consequently, the aim of our study is to explore claims that take the form of social-media posts, particularly those reflecting people’s negative reactions to VAW, which is the type of social problem to be detected.

3- Large and small pre-trained language models

Recent advances in transformer-based language models pre-trained on vast quantities of text have contributed to the emergence of large language models (LLMs) such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), and PaLM (Chowdhery et al., 2023), among many others. Indeed, the superior generative capacity of LLMs has led to a revolutionary breakthrough in deep learning and natural language processing. To comprehend and produce natural language, LLMs employ complex neural networks containing billions of parameters that are frequently trained on large corpora from the web using a self-supervised learning approach to discover language patterns and subtleties (Raiaan et al., 2024). LLMs have significantly enhanced their language under-

standing and text generation capabilities and, consequently, extended their applicability to a wider range of natural language processing tasks, e.g. emotion recognition, question answering, sentiment analysis, summarisation, text analysis, text classification, text generation, and translation, just to name a few.² Compared to traditional machine-learning and deep-learning models, the advantage of LLMs lies in the fact that, once they are trained, they can be reused to perform tasks for which they hadn't been explicitly trained, thus contributing to resource adaptability. Moreover, LLMs can achieve accurate results without the need to adopt the traditional model for text classification (i.e. data pre-processing, feature extraction, and classifier training), thus contributing to resource efficiency.

In the field of LLMs, OpenAI has made substantial progress in developing models that exhibit remarkable proficiency in generating high-quality human-like responses. For example, ChatGPT is currently based on GPT-3.5 and GPT-4, where the former has been trained with 175 billion parameters, and the latter has 1.8 trillion parameters (i.e. over 10 times larger than GPT-3.5). It is assumed that the ability of LLMs to comprehend intricate language patterns and produce appropriate human-like responses is critically affected by the number of parameters with which the models were trained. In other words, more parameters mean a more powerful LLM. Therefore, the more parameters an LLM has, the better it can learn when prompted. However, ChatGPT poses two problematic issues for researchers. On the one hand, ChatGPT is constantly updated, so different versions of the model can yield different results. Consequently, this issue raises two serious problems over time: (a) researchers do not have control over the performance of the model, and (b) they can't replicate the results of their experiments. On the other hand, OpenAI does not provide free API services for GPT-3.5

2. Minaee et al. (2024) presented an overview of the state of the art of LLMs, focusing on three popular LLM families (i.e. GPT, LLaMA, and PaLM) and other representative models.

and GPT-4 models. Therefore, the cost of using ChatGPT in a production environment can be prohibitive if vast amounts of data have to be processed. Alternatively, using open-source LLMs is also not a viable solution for most researchers because of the demanding hardware requirements. For example, Llama 3.1, with 405 billion parameters, cannot be locally run on consumer-grade computers but on hardware from specialised AI research labs, high-performance computing centres, and cloud computing providers. For these reasons, Minaee et al. (2024) explained that, although LLMs can get better accuracy and performance, there is a current research trend to develop smaller models as a cost-effective alternative to use on specific tasks that do not require the capabilities of LLMs. For this reason, we experimented with a smaller but stable pre-trained language model, which does not undergo changes or updates and can be easily integrated into downstream applications.

For our experiment, we employed Microsoft Orca 2 7b³ to detect the social problem of VAW. Table 1 shows some technical specifications of this SLM.

Table 1. Specifications of Microsoft Orca 2 7b.

Type	Parameters	File size	RAM	Compression
LlaMA 2	7 billion	3.74 GB	6.33 GB	4-bit quantisation (Q4.0)

We chose Microsoft Orca 2 because this instruction-based model is suitable as a reasoning engine over knowledge provided to the model through the prompt (i.e. in-context instructions). Mitra et al. (2023) demonstrated that it can outperform language models of similar size and achieve a performance level comparable to models that are five to ten times larger, especially on zero-shot tasks.

3. <https://huggingface.co/TheBloke/Orca-2-7B-GGUF>

4- Methodology

4.1. Overview

The goal of this study is to discover the most efficient unsupervised classification method (i.e. based on in-context learning or on synthetic corpora) for leveraging the SLM to successfully perform the intended task. In the case of in-context learning, we explored the prompting technique of zero-shot with and without chain-of-thought. Zero-shot prompting involves the model being fine-tuned only with a direct specification of the task. In chain-of-thought prompting, the model is guided to explain the reasoning that led it to produce a particular answer.

4.2. Classification based on in-context learning

In the first step, we performed text classification with zero-shot prompting (with and without chain-of-thought). We employed the SLM with hyperparameter settings that could reduce creativity in the output (i.e. Temperature: 0.1, Top K: 1, and Top P: 0.1). The maximum number of tokens was 200, as we wanted to avoid generating long responses that could be irrelevant. Moreover, we experimented with two components to be included in the prompt: the natural language description of the problem type (D) and the request for step-by-step reasoning (R). As prompts can go with (+) or without (-) these components, we generated four prompt templates, which we called +D+R, -D+R, +D-R and -D-R. To illustrate, here is the +D+R prompt template, where the placeholder [*language*] was replaced with *Spanish* or *English*, and [*user-generated content*] was filled in with one tweet from the corresponding test dataset:

You are an AI assistant that has been trained to identify [*language*] tweets that describe some community problems in the city. Community problems are expressed through statements that show a negative reaction to a current situation that affects citizens' quality of life. You are an expert on the community problem "violence against women", which is when "a woman suffers violence, maltreatment, or threats". Here is the [*language*] tweet: "[user-generated

nerated content]”. Does the tweet describe the community problem “violence against women”? Generate a JSON object with the fields “Answer” and “Explanation”. The Answer value must be “Yes” or “No” for the above question. The Explanation value is text that explains your step-by-step reasoning.

JSON object:

4.3. Classification based on synthetic corpora

In the second step, we performed similarity-based text classification through a procedure that we called the STECS method (i.e. SOM-based TExt Classification with Synthetic corpora). First, we employed the SLM to construct a corpus of synthetically generated texts about the problem type. We constructed a prompt template where placeholders were programmatically filled in as follows: *[keywords]* with a comma-separated sequence of keywords from Table 2, *[style]* with *formal*, *informal*, *colloquial* or *slang*, and *[speech_function]* with *criticism* or *complaint*. Here is the prompt template that allows generating various customised prompts:

You are a citizen who publishes posts on social networks about community problems in your city. Write one post between 25 and 36 words in *[language]* as a *[speech_function]* about the community problem “violence against women”, which is described with the following keywords: “*[keywords]*”. Write the post in *[style]* style. Your response must only include the post. Remember, your post should fall within the word count range specified above.

Post:

Table 2. Keywords of VAW.

Spanish	English
amenaza, mujer	threat, woman
abuso, mujer	abuse, woman
maltrato, mujer	maltreatment, woman
maltrato, amenaza, mujer	maltreatment, threat, woman
violencia, mujer	violence, woman
violencia, amenaza, mujer	violence, threat, woman
violencia, maltrato, mujer	violence, maltreatment, woman
violencia, abuso, mujer	violence, abuse, woman
abuso, amenaza, violencia, mujer	abuse, threat, violence, woman
amenaza, maltrato, mujer, violencia	threat, maltreatment, woman, violence

Each synthetic corpus consisted of 720 posts, where nine different posts were generated for each of the ten lexical contexts in each of the four styles and each of the two speech functions (i.e. $9 \times 10 \times 4 \times 2 = 720$). We tweaked settings such as Temperature, Top K, and Top P to increase diversity and creativity in the output (i.e. Temperature: 0.9, Top K: 50, and Top P: 0.9).

Second, we converted each tweet in the test dataset and each text in the synthetic corpus into an embedding generated by Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019), a fine-tuned BERT model (Devlin et al., 2019) that uses Siamese and triplet network structures to derive semantically meaningful text embeddings. For our experiment, we used S-BERT with the multilingual pre-trained model *distiluse-base-multilingual-cased-v1*, which supports languages such as English and Spanish, among many others (Reimers and Gurevych, 2020).

Third, a score was computed for each tweet in the test dataset based on the similarity between the embedding of the tweet and the embeddings of the synthetic corpus. To obtain the similarity score, we trained a self-organising map (SOM) (Kohonen, 2001) with the embeddings of the synthetic corpus, passed the embedding of each tweet through the SOM, and used the Gaussian kernel ($\sigma = 1$) as the measure to obtain the similarity score between each tweet and its most significant Best Matching

Unit (BMU) in the SOM, that is, the unit in the SOM with the smallest distance to the input embedding. A SOM is a (typically, two-dimensional) single-layer neural network used for reducing the dimensionality of high-dimensional datasets but preserving the topological structure of the data. The hyperparameter settings of the trained SOM were Width: 12, Height: 12, Learning rate: 0.5, Learning radius: 6, and Iterations: 8000.

Fourth, we considered each similarity score in the test dataset as a potential cut-off value, where scores above the value were treated as positive and those below the value as negative, thus calculating true positives, true negatives, false positives, and false negatives for each cut-off.

Fifth, we evaluated each cut-off value and then selected the optimal threshold based on the value that provided the best F1 score.

4.4. Evaluation measures

We evaluated the performance of each step with measures such as Precision, Recall, and F1. With respect to binary classification, it should be recalled that most of the evaluation measures are computed over a 2x2 contingency matrix—as shown in Table 3, where TP, FP, FN and TN denote the number of true positives, false positives, false negatives, and true negatives, respectively.

Table 3. Contingency matrix for binary classification.

			Expected	
			Does the tweet really pertain to class x ?	
			yes	no
			Predicted	Was the tweet classified as x ?
	no	FN		TN

In this regard, we employed two typical evaluation measures that come from information retrieval, i.e. Precision and Recall, which are shown in Equations (1) and (2).

(1)

$$Precision = \frac{TP}{TP + FP}$$

(2)

$$Recall = \frac{TP}{TP + FN}$$

Moreover, we used one of the most popular measures that combines Precision and Recall, i.e. F1, which is presented in Equation (3).

(3)

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

5- Experiment

5.1. Test datasets

We employed two small test datasets of Spanish and English tweets that resulted from queries based on a few seed words that are closely related to VAW. Indeed, both datasets were derived from a corpus of tweets on VAW that had been previously collected for a different study.⁴ Table 4 shows the details of the test datasets, where + and – represent the number of tweets that do and do not describe the target problem type, respectively.

4. Further information about this corpus is given in Alameda Hernández (2024).

Table 4. Test datasets.

Language	Tweets	+	–	Seed words
Spanish	177	89	88	violencia contra mujer violencia de género maltrato mujer
English	110	55	55	violence against women gender-based violence domestic violence

To illustrate, we present some positive (Examples 4 and 5) and negative (Examples 6 and 7) tweets in Spanish and English:

- (4) Como mujer Tigrense me siento profundamente ofendida por el maltrato y violencia de Ernesto Paraqueima para con las mujeres de la ciudad. Reafirmar estereotipos y burlarse no solo de nosotras, sino de otras personas es inaceptable
- (5) 300 women are raped every day in the UK, and only 3 of those cases will make it to court. As a survivor of male violence and abuse, I'd like the Home Secretary to focus on dealing with ALL perpetrators of violence and abuse against women, not on demonising communities.
- (6) Gerardo Herrero estrena #BajoTerapia, una historia entre la comedia dramática y el juego psicológico sobre los roles de género, las dinámicas de pareja y la violencia contra las mujeres
- (7) Are you a refugee woman or girl or do you know any refugee women or girls experiencing gender based violence? We are here to help you.

Whereas Examples (4) and (5) clearly frame mistreatment and violence faced by women as an ongoing social issue that affects their communities, Examples (6) and (7) do not represent claims reflecting social concern about this problem type. For instance, Example (7) represents an offer of help with a service-oriented goal rather than a call to awareness or action regarding a societal problem.

Two annotators labelled all the tweets in the above test datasets, where a third annotator made the final decision only in case of disagreement. Annotators were provided with the following written instructions to categorise the tweets for this problem type, together with ten positively and ten negatively labelled tweets and the reasoning behind each annotation:

You should read the tweet carefully and decide whether or not the author of the tweet thinks that “violence against women (i.e. women who suffer violence, maltreatment, or threats)” is a current community problem. If the author thinks that the problem occurs, categorise the tweet as positive; otherwise, categorise it as negative. If a tweet doesn’t provide enough information to determine that the above community problem occurs, then categorise it as negative. You should keep in mind that community problems are public beliefs expressed through claims that show a negative reaction to a current situation that affects citizens’ quality of life in a given community. Therefore, (a) community problems are beliefs, so such subjective judgments and thoughts are taken by the individual to be true, although they may or may not actually be true, and (b) community problems are about present situations, so they exist at the present moment, but their existence may also stretch into the past and the future.

5.2. Results

First, the performance of the SLM was evaluated with zero-shot prompting based on two different settings: whether or not the prompt included the description of the problem type (+D/-D), and whether or not the prompt included the request for step-by-step reasoning (+R/-R). Second, we performed similarity-based classification with the STECS methods; the optimal thresholds for the Spanish and English test datasets were 0.688 and 0.672, respectively, based on the cut-off value that provided the best F1 scores. The results of the experiment are presented in Table 5 and Table 6.

Table 5. Performance of Microsoft Orca 2 7b with the Spanish test dataset.

Approach	Precision	Recall	F1
zero-shot (+D+R)	0.716	0.652	0.682
zero-shot (-D+R)	0.644	0.730	0.684
zero-shot (+D-R)	0.593	0.719	0.650
zero-shot (-D-R)	0.576	0.899	0.702
STECS	0.830	0.876	0.852

Table 6. Performance of Microsoft Orca 2 7b with the English test dataset.

Approach	Precision	Recall	F1
zero-shot (+D+R)	0.725	0.673	0.698
zero-shot (-D+R)	0.651	0.745	0.695
zero-shot (+D-R)	0.589	0.964	0.731
zero-shot (-D-R)	0.548	0.927	0.689
STECS	0.754	0.891	0.817

5.3. Discussion

On the one hand, no specific prompt setting in zero-shot prompting seemed the most effective. For example, whereas the best and worst F1 scores with the Spanish test dataset were achieved with the -D-R (0.702) and +D-R (0.65) settings, respectively, +D-R (0.731) was the best and -D-R (0.689) was the worst with the English test dataset.

On the other hand, the similarity-based classification method outperformed all the others in Precision and F1 scores. Therefore, although Microsoft Orca 2 was developed to enhance the ability of an SLM to reason in zero-shot settings (Mitra et al., 2023), the STECS method may be regarded as a better choice. Moreover, considering that inference time and memory usage are relatively higher with SLMs in in-context learning, SOMs in the STECS method can help solve this problem when the system is deployed in high-throughput environments for real-time processing with resource-constrained devices.

6- Conclusions

VAW can be analysed through social computing by examining social-media posts through computational methods to identify recurring patterns (e.g. about the different types of violence in VAW) and monitor emerging trends (e.g. shifts in public opinion), thus allowing researchers, policymakers, and organisations to gain real-time, data-driven insights into this social problem. In this context of social sensing, we explored two strategies for using an SLM, specifically Microsoft Orca 2 7b, to identify Spanish and English tweets related to VAW: (a) in-context learning and (b) creating a synthetic corpus to train an unsupervised classifier. SLMs were preferred over LLMs because their fewer computational requirements make them easier to implement in real-world applications.

For in-context learning, we experimented with zero-shot prompting, with and without chain-of-thought reasoning. Our results showed that the performance of the SLM is highly sensitive to factors such as the structure of the prompt (e.g. whether or not it includes a description of the problem) and the language used. We did not find conclusive evidence pointing to one prompting technique as consistently superior for optimising results.

In contrast, we demonstrated that training SOMs with embeddings generated from synthetic corpora for text classification (i.e. the STECS method) was the most effective approach. This unsupervised method relies on the similarity score, calculated with Gaussian kernel as the distance-based measure, used to match the embedding of an input to its closest prototype vector (or BMU) in the SOM. It should be noted that one of the parameters that should be predefined in this method is the similarity threshold α : any text with a similarity score above α is classified as positive, while those falling below it are labelled as negative. However, the difference between the values of α for the Spanish and English test datasets is so small (i.e. 0.016) that simply choosing the mid-value can ensure a narrow margin of error for the optimal threshold. Future research should examine whether the STECS method can provide a strong foundation for systems designed to detect many other community problems (e.g. institutional elder abuse, black economy, etc.).

Acknowledgements

This publication is part of the R&D&I project PID2023-147137NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU, and the research project CIPROM/2023/29, funded by “Direcció General de Ciència i Investigació” Generalitat Valenciana (Spain).

References

- Alameda Hernández, A. (2024). “Social media detection of texts on the social problem of violence against women within the multimodal intelligent system ALLEGRO”. *Approaches to knowledge representation and language*. Comares, 63-76.
- Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). “Social sensing of floods in the UK”. *PLoS ONE*, 13 (1). <https://doi.org/10.1371/journal.pone.0189327>
- Berger, P. & Luckman, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Penguin Education.
- Best, J. (1989). “Dark figures and child victims: Statistical claims about missing children”. *Images of issues*. Aldine de Gruyter, 21-25.
- Best, J. (2017). *Social problems*, 3rd edition. W.W. Norton & Company.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., & Fiedel, N. (2023). “PaLM: Scaling language modeling with pathways”. *Journal of Machine Learning Research*, 24, 1-113.
- Council of Europe (2021). *What is gender-based violence?* <https://www.coe.int/en/web/gender-matters/what-is-gender-based-violence>. Retrieved on 12 June 2025.

- Davis, K. (1937). "The sociology of prostitution". *American Sociological Review*, 2 (5), 746-55. <https://doi.org/10.2307/2083827>
- DeFronzo, J. & Gill, J. (2020). *Social problems and social movements*. Rowman & Littlefield.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.
- Endsley, M. R. (2000). "Theoretical underpinnings of situation awareness: A critical review". *Situation awareness analysis and measurement*. Lawrence Erlbaum Associates, 3-32. <https://doi.org/10.1201/b12461>
- Holstein, J. & Miller, G. (2003). *Challenges & choices*. Aldine de Gruyter.
- Kohonen, T. (2001). *Self-organizing maps*, 3rd ed. Springer. <https://doi.org/10.1007/978-3-642-56927-2>
- Letukas, L. (2014). *Primetime pundits: How cable news covers social issues*. Lexington Books.
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large language models: A survey*. <https://doi.org/10.48550/arXiv.2402.06196>. Retrieved on 13 April 2025.
- Mitra, A., Del Corro, L., Mahajan, S., Cudas, A., Simoes, C., Agarwal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., Palangi, H., Zheng, G., Rosset, C., Khanpour, H., & Awadallah, A. (2023). *Orca 2: Teaching small language models how to reason*. <https://doi.org/10.48550/arXiv.2311.11045>. Retrieved on 25 May 2025.
- Mooney, L. A., Clever, M., & Van Willigen, M. (2021). *Understanding social problems*. Cengage.
- Musto, C., Semeraro, G., Lops, P., & De Gemmis, M. (2015). "CrowdPulse: A framework for real-time semantic analysis of social streams". *Information Systems*, 54, 127-146. <https://doi.org/10.1016/j.is.2015.06.007>
- OpenAI (2023). *GPT-4 technical report*. <https://doi.org/10.48550/arXiv.2303.08774>. Retrieved on 24 January 2025.
- Periñán-Pascual, Carlos (2023) "From Smart City to Smart Society: A quality-of-life ontological model for problem detection from user-generated content". *Applied Ontology*, 18 (3), 263-306. <https://doi.org/10.3233/AO-230281>

- Raiaan, M. A. K., Mukta, Md S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). "A review on large language models: Architectures, applications, taxonomies, open issues and challenges". *IEEE Access*, 12, 26839-26874. <https://doi.org/10.1109/ACCESS.2024.3365742>
- Reimers, N. & Gurevych, I. (2019). "Sentence-BERT: Sentence embeddings using Siamese BERT networks". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982-3992. <https://doi.org/10.18653/v1/D19-1410>
- Reimers, N. & Gurevych, I. (2020). "Making monolingual sentence embeddings multilingual using knowledge distillation". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4512-4525. <https://doi.org/10.18653/v1/2020.emnlp-main.365>
- Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). "Tweet analysis for real-time event detection and earthquake reporting system development". *IEEE Transactions on Knowledge and Data Engineering*, 25 (4), 919-931. <https://doi.org/10.1109/tkde.2012.29>
- Shepherd, L. (2008). *Gender, violence, and security: Discourse as practice*. Zed books. <https://doi.org/10.5040/9781350220317>
- Spector, M. & Kitsuse, J. (1977). *Constructing social problems*. Transaction Publishers.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and efficient foundation language models*. <https://doi.org/10.48550/arXiv.2302.13971>. Retrieved on 3 February 2025.
- United Nations (1993). *Declaration on the elimination of violence against women*. https://www.un.org/en/genocideprevention/documents/atrocities-crimes/Doc.21_declaration%20elimination%20vaw.pdf. Retrieved on 27 February 2025.