The Roe v. Wade sentence: an analysis of tweets trough Symmetric Non-Negative Matrix Factorization

Maria Gabriella Grassia[a], Marina Marino[a], Rocco Mazza[b], Agostino Stavolo[a]

[a] Department of Social Sciences, University of Naples "Federico II", Naples, Italy;
mariagabriella.grassia@unina.it; marina.marino@unina.it;
agostino.stavolo@unina.it
[b] Department of Political Science, University of Bari "Aldo Moro", Bari, Italy;
rocco.mazza@uniba.it

Abstract

In recent years, social media has become the main field for sourcing textual data. In particular, microblogging platforms such as Twitter make it possible to study users' online discussions by understanding and analysing the opinions, comments, and experiences that users share on different issues. One of the most debated issues in recent years is the voluntary termination of pregnancy. In particular, the United States Supreme Court's Roe v. Wade ruling has brought the abortion debate back to the forefront. Indeed, after the annulment of the ruling that restricted the right to abortion in the U.S., the response of online users has been crucial. Indeed, the aim of the paper is to understand what the major topics of discussion have been in the wake of the ruling. To do this, semantic clusters of terms were created through a symmetrical matrix reduction technique, Symmetric Non-Negative Matrix factorization.

Key words: Twitter, Symmetric non-negative matrix factorization, Lexical matrix decomposition

1. Introduction

On 24th June 2022, the U.S. Supreme Court overturned the sentence Roe v. Wade, which established the constitutional right to abortion in the United States in 1973. The Republican-appointed justices voted to strike down the federal right, while the other Democratic justices voted against it.

With the annulment of the ruling, individual U.S. states now have the power to establish their own laws regarding the right, or not, of a woman to have a termination of pregnancy, which Roe v. Wade allowed by the 24th week. In the absence of a law from Congress regulating abortion at the federal level, it will mean that each state can decide whether to allow abortions, whether to ban them always or under certain circumstances.

This event produced legal-normative consequences regarding whether or not voluntary termination could be accessed, but also social consequences in that it opened up a wide debate about the freedom to choose and to protect women's bodies. In addition, people seeking abortions will have to move from the city where they live and travel to more distant clinics or hospitals that allow the practice.

The outcome of the ruling has created a wide online debate among citizens who have expressed positions for or against the incident. This is because social networks and microblogging platforms are used as tools of political exchange: the debate on microblogging platforms has been studied before (Graells-Garrido et al. 2020; Sharma et al., 2017; Zang et al. 2016). Data from social media platforms such as Twitter offer new possibilities to study these dynamics because it provides a public arena for information gathering and opinion formation (Grassia et al. 2022).

So, the following work contributes to the study of the online debate on the issue of abortion through the analysis of tweets posted by users following the ruling. To understand the issues that users focused on in the online discussion, we used Symmetric Non-Negative matrix factorization (symNMF), which is a symmetric matrix reduction method used in clustering operations to create semantic clusters of terms.

2. Symmetric Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF) is an unsupervised matrix decomposition method that decomposes a matrix $\mathbf{X} \in R^{n \times m}$ by a product of two factor matrices $\mathbf{X} \approx \mathbf{WH}$, where $\mathbf{W} \in R^{n \times k}$ and $\mathbf{H} \in R^{k \times m}$, both with nonnegative elements. The product WH is an approximate factorization of rank at most k that it is assumed to satisfy the condition $k \ll min\ \{m,n\}$ (Gaujoux and Seoighe, 2010). The value of the parameter k shows the numbers of factors to be used to explain data (Casalino et al. 2016).

The NMF has the advantage that provides easily interpretable results because the extracted latent features are parts of the original data. Recently, NMF is widely used in text mining and document clustering but also for topic modeling in short texts. However, NMF is not a general clustering method that can be applied in every condition, as it is dependent on the linear or non-linear structure of clusters. (Kuang et al. 2015).

To solve the problem, it is used the symmetric version of NMF, the Symmetric Non-Negative Matrix Factorization (symNMF). This method factorizes a symmetric matrix input $\mathbf{A}$ by the product of two matrices $\mathbf{A} \approx \mathbf{HH^T}$, where $\mathbf{H}$ is the cluster assignment and its transpose $\mathbf{H^T}$ (Jia et al. 2021). Specifically, the matrix $\mathbf{H}$ is a nonnegative matrix of size $n \times k$, and k is the number of clusters requested.

Suppose the data points of the same group have high similarity values and the data points of the different groups have weak similarity values. So, a better approximation of $\mathbf{A}$ defines the cluster structure because the largest entry in the i-th row of $\mathbf{H}$ indicates the clustering assignment of the i-th data point, according to the nonnegativity of $\mathbf{H}$. The usual approach for approximating the input matrix $\mathbf{A}$ is to minimize the Frobenius norm $min_{H \geq 0} ||A \approx HH^T||^2_F$, and it can be related to a generalized form of many clustering objectives (Kuang et al. 2012).

SymNMF has been shown to be more effective for nonlinearly separable data than NMF (Kuang et al. 2015). It is important to know that symNMF is equal to spectral clustering, but Vangara et al. (2021) showed that it assumes better performance than k-means and spectral clustering. Kabir et al. (2020) stated that symNMF is more functional for clustering because it converts spectral clustering into an optimization problem with stationary point solutions. It has also been used as a graph clustering method (Luo et al.2021) and for topic extraction from a lexical matrix (Grassia et al. 2022; Yan et al., 2013)

3. Methodology

According to the open access academy API, we extracted English-language tweets using #RoeVsWade in the week following the outcome of the ruling (June 24-30, 2022). The volume of data extracted was very large: 943,696 tweets were extracted. In this regard, we eliminated retweets and identified 214,469 documents.

As part of an information mining process, it is necessary to process texts and obtain a set of structured data that can be elaborated using statistical techniques. So, several text pre-processing operations were carried out to transform the textual data (i.e., tweets) into structured data.

Documents are parsed and tokenized, resulting in a set of distinct strings (*tokens*) separated by blanks, punctuation marks, or other types of special characters (e.g., hashtags). These tokens correspond to the terms used in the vocabulary. The particular scheme achieved by tokenization is commonly known as bag-of-words (BoW), as it treats each document as a multiset of its tokens, without regard to grammatical and syntactic roles.

Once documents have been atomized into their basic components, pre-processing is necessary to reduce linguistic variability (Uysal and Gunal, 2014). First, all characters of the terms were changed to lowercase. To account for language variety, the following were carried out other *normalization* operations, such as correcting misspelled terms or removing numbers (Misuraca and Spano, 2018).

To reduce morphological variability, *lemmatization* was carried out, where each term was returned to its canonical form (verbs are returned to the present infinitive, nouns, and adjectives to the masculine singular). When texts have been pre-processed, it is possible to construct the so-called *vocabulary* by stacking identical terms and counting the number of occurrences of each vocabulary (type) in the document collection. To avoid uninformative terms, the vocabulary can be trimmed by removing so-called *stop-words*, i.e., the common terms used in the specific language and domain analysed (prepositions, conjunctions, etc.). For the same reason, rare terms with a low number of occurrences are usually removed from the vocabulary.

These phases returned a database composed of 2.819,642 tokens, 63,150 types and 214,469 documents. In the final stage of the pre-processing process, we applied the matrix vector space model of documents and words. Each document can be seen as a vector in *p*-dimensional vector space spanned by the terms belonging to the vocabulary. We created a term-document matrix, where *term frequency* is used to express the relative importance of each term in each document. Raw frequency weights are calculated as the number of occurrences of a term in a document and correspond to the absolute frequency. This system of weights results in the creation of a sparse matrix, which, in the case of NMF, leads to numerous problems with the interpretability of the result.

According to Yan et al. (2013), to reduce the sparsity of the term-document matrix, we transform it into a *co-occurrence* matrix, that represents the number of times two terms w_i and w_j co-occur together. For each pair of vectors, we define the *cosine* similarity measure that expresses the association of terms. Cosine similarity measures the similarity between two vectors of an inner product space. It is defined by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. Thus, we created a similarity matrix, and we applied the symNMF to the similarity matrix for defining semantic clusters of terms.

4. Preliminary results

Specifically, we define four semantic clusters.

Table 1: Semantic clusters by SymNMF

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Odd	Trample	Hypocrisy	Healthy
Prevent	Witness	Demand	Daughter
Process	Terrible	Duty	Disappoint
Progress	Upheld	Moral	Embarrassment
Miscarriage	Trash	Legalize	Disregard
Oppose	Violation	Fail	Harm
Overturn	Stripping	Democracy	Freedom
Practice	Norm	Jesus	Mother
Respect	Rights	Life	Children
Vote	Amendment	Church	Violence

From the analysis of the clusters extracted from the matrix, we noticed how there was a strong aversion to the outcome of the judgment. Table 1 shows the top ten terms associated with the clusters created.

We can see that the first cluster focuses on the narrative of the ruling and the reversal of the right to abortion. The Supreme Court overturned the ruling legalizing abortion, reaffirming that authority now reverts to the people and state representatives. All of this led to a strong outrage from Twitter users against the ruling, highlighting how the result will lead to negative consequences for the right to abortion. Through the overturning of the ruling numerous women's rights are going to be violated, primarily the freedom to choose one's own body and choose whether to continue the pregnancy. This is also interpreted in the second cluster, which highlights a number of terms such as 'trample', 'trash' and 'terrible' which indicate a strong aversion of online users to the decision taken.

Added to this are the users' criticism of the Catholic world and the church, pointing out how the church is hypocritical in not recognizing individuals who decide to have an abortion full legality. Abortion presents a profound moral issue, and the church and bishops have called the incident a "historic day" as defenders of the so-called "pro-life" movements. The rationale behind pro-life movements is to equate the gynaecological practice with the killing of an embryo, which acquires the status of a person at the moment of conception, and to consider abortion a highly damaging intervention for health.

The latest cluster created, however, shows parents' concern for their children's future. Users wonder what will happen after the ruling and hope that their children will be able to grow up in a state that

guarantees more rights than they had. And it is especially the issue of violence that creates discontent, as people are frightened and perplexed about the future of girls should they experience violence and have difficulty accessing abortion practices.

5. Conclusions

Nowadays, short texts have become an interesting form of text information, such as social media posts, question titles, and comments on posts. Short texts from the Internet are often extremely short, noisy, and ambiguous, imposing great challenges to clustering. The biggest difficulty is that each short text only holds very few word tokens.

This preliminary work aims to show how symNMF is an optimal technique for creating term clusters. About what has been analysed, it has been shown how Twitter users had a strong reaction to the overturning of the U.S. Roe v. Wade ruling by expressing strong criticism of the Supreme Court's choice.

For future developments, we are improving the level of efficiency in the construction of clusters by automatically defining the number of k clusters to be detected and adopting of specific metrics that effectively measure the quality of clusters.

References

[1] Casalino, G., Del Buono, N., Mencar, C.: Non-negative matrix factorizations for intelligent data analysis. Non-negative Matrix Factorization Techniques (pp. 49-74), Springer, Berlin, (2016).

[2] Gaujoux, R., Seoighe, C.: A flexible R package for nonnegative matrix factorization. BMC bioinformatics, 11(1), 1-9, (2010).

[3] Graells-Garrido, E., Baeza-Yates, R., Lalmas, M.: Representativeness of abortion legislation debate on twitter: A case study in Argentina and Chile. In Companion Proceedings of the Web Conference 2020 pp. 765-774, (2020).

[4] Grassia M.G., Marino M., Mazza R., Stavolo A.: Analysis of the public debate on DDL Zan on Twitter: an application of the Structural Topic Model in Proceedings of the 16th International Conference on Statistical Analysis of Textual Data, Vol. 1, pp. 67-73, (2022).

[5] Grassia M.G., Marino M., Mazza R., Misuraca M., Stavolo A.,: Topic modeling for analyzing the Russian propaganda in the conflict with Ukraine in Book of Short Papers of the ASA Conference 2022 - Data-Drive Decision Making, *In press*, (2022)

[6] Jia, Y., Liu, H., Hou, J., Kwong, S., Zhang, Q.: Self-supervised symmetric nonnegative matrix factorization. IEEE Transactions on Circuits and Systems for Video Technology (2021)

[7] Kuang, D., Ding, C., Park, H.: Symmetric nonnegative matrix factorization for graph clustering. In Proceedings of the 2012 SIAM international conference on data mining (pp. 106-117). Society for Industrial and Applied Mathematics, (2012).

[8] Kuang, D., Choo, J., Park, H.: Nonnegative matrix factorization for interactive topic modeling and document clustering, in Partitional Clustering Algorithms pp. 215-243. Springer, Cham (2015).

[9] Luo, X., Liu, Z., Jin, L., Zhou, Y., Zhou, M.: Symmetric nonnegative matrix factorization-based community detection models and their convergence analysis, IEEE Transactions on Neural Networks and Learning Systems (2021)

[10] Misuraca, M., Spano, M.: Unsupervised analytic strategies to explore large document collections. In International Conference on the Statistical Analysis of Textual Data, pp. 17-28. Springer, (2018).

[11] Sharma, E., Saha, K., Ernala, S. K., Ghoshal, S., De Choudhury, M.: Analyzing ideological discourse on social media: A case study of the abortion debate. In Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas, (2017)

[12] Uysal A, Gunal S: The impact of preprocessing on text classification. Information Processing and Management 50(1):104–112, (2014)

[13] Vangara, R., Rasmussen, K. Ø., Chennupati, G., Alexandrov, B.: Determination of the number of clusters by symmetric non-negative matrix factorization. In Big Data III: Learning, Analytics, and Applications Vol. 11730, pp. 104-113. SPIE, (2021)

[14] Yan, X., Guo, J., Liu, S., Cheng, X., Wang, Y. Learning topics in short texts by non-negative matrix factorization on term correlation matrix, in Proceedings of the 2013 SIAM International Conference on Data Mining, pp. 749-757, Society for Industrial and Applied Mathematics, (2013).

[15] Zhang, A. X., Counts, S.: Gender and ideology in the spread of anti-abortion policy. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems pp. 3378-3389, (2016).