Statistical Journal of the IAOS 40 (2024) 535–548
DOI 10.3233/SJI-240017
IOS Press

Integrating word embedding and topic modeling for sentiment analysis: A case study on the social mood on economy

Mauro Bruno*, Elena Catanese and Massimo De Cubellis
Italian National Institute of Statistics (Istituto Nazionale di Statistica – ISTAT), Rome, Italy

Abstract. In recent years, textual analysis and embedding spaces have become essential and complementary tools for sentiment analysis in National Statistics Institutes' research, owing to their ability to summarize discussed topics effectively. Istat has developed an innovative tool, wordembox, which allows external users to explore the outputs of popular word embedding algorithms, such as Word2Vec and FastText. This tool enriches the analysis with a novel graph functionality, enabling users to discover clusters of words and facilitating implicit topic modeling.
This article focuses on Social Mood on Economy (SME) posts over a period in which the index recorded a strong downward trend: the first month of the Russia-Ukraine conflict at the beginning of 2022. We compare findings from wordembox with standard topic modeling techniques, including Bayesian Latent Dirichlet Allocation (LDA), Top2Vec, and BERTopic, recent methods that extract clusters from word embedding spaces. These techniques show coherent results, and their combined use in textual analysis may create a synergy that enhances the informative content of synthetic indexes such as 'Social Mood on Economy Index (SMEI)'.

Keywords: Word embedding, topic modeling, natural language processing, social media

1. Introduction

Within the domain of unsupervised machine learning (ML), word embeddings (WEs) are gaining prominence. These methods efficiently derive word representations from vast, unstructured textual data without supervision. Generally, WE representations capture syntactic and semantic consistencies found in language patterns [1,2]. Every relationship appears as a distinct vector offset, facilitating reasoning based on vector orientation. The fundamental principle behind these vector representations of words is encapsulated in the "distributional hypothesis," which posits that the essence of a word can be understood by examining its adjacent words [3].

The Italian National Institute of Statistics (Istat) developed in 2018 an open-source tool (wordembox) to analyze the content of posts collected since 2016 from the X platform, formerly known as Twitter. Employing word embedding algorithms trained on X posts, as a corpus of short texts, wordembox facilitates the exploration of embedding spaces through two-dimensional graphical representations. In these graphs, words are displayed as nodes, and the connections between them, represented by edges, indicate semantic or syntactic relationships. This paper employs wordembox to examine the word embedding space associated with key terms relevant to the Russia-Ukraine conflict. It focuses on understanding the evolution of their representation over time and highlights the differences in word embedding models trained on varying sets of posts. A detailed analysis of the wordembox is out of the scope of the present work. However, Annex 1 provides an overview of the architecture and main functionalities.

Additionally, unsupervised ML techniques are very promising in highlighting emerging topics from posts related to certain phenomena. Analyses of "most discussed topics" require natural language processing techniques to discover semantic patterns. Generally, this can

*Corresponding author: E-mail: mbruno@istat.it.

1874-7655/$35.00 © 2024 – IOS Press. All rights reserved.

be done via either Topic Models, Bayesian statistical models, or WE models derived from Neural Networks. These two approaches differ substantially in how they handle word representation. While in the WE model, each word (usually) has a unique qualitative representation, in topic models, each word has a quantitative representation associated with a probability of belonging to a specific cluster, estimating the expected absolute frequencies of a word. The present paper compares these methods, focusing on the Russia-Ukraine conflict.

The structure of this paper is outlined as follows: Section 2 provides an overview of research projects that use word embedding techniques to support statistical analysis both in the context of official statistics and academia. Section 3 offers a comprehensive overview of the models employed in our simulations, specifically focusing on Word2Vec, LDA, Top2Vec, and BERTopic. In Section 4, we compare a manual approach using wordembox and automatic clustering methods like Biterm, Top2Vec, and BERTopic, highlighting key findings, similarities, and differences. In Section 5, we conclude and discuss potential future enhancements using new and different methods that allow importing previously trained Word Embedding Spaces. Annex 1 delves into the architecture of wordembox, detailing its core functionalities that facilitate the exploration of embedding spaces.

2. Context

Word embedding models can be very useful as tools to support statistical analysis. This paragraph will describe some use cases from official statistics where these models have been implemented.

Istat has published a paper [4] that provides a detailed analysis of sentiment analysis and word embedding models applied to X data on climate change in Italy. The Social Mood on Climate Change Index is calculated using an unsupervised lexicon-based approach. Such an approach is currently used in Istat to calculate the Social Mood on Economy Index.[1] The daily pipeline consists of three fundamental steps: (i) collection and pre-processing (text cleaning) of the daily sample of public posts related to climate change; (ii) estimate of the sentiment score of each post in the sample; (iii) calculation of the daily value of the index as a synthesis of the sentiment values of the entire sample. The approach used in this work demonstrates that state-of-the-art ML methods significantly improve the accuracy and relevance of sentiment analysis and topic modeling compared to traditional methods based solely on keywords and frequency counts.

Some authors apply innovative NLP techniques, such as BERT, in a study of consumer confidence dynamics through online news analysis [5]. Using a dataset of over 1.8 million articles, the analysis employed advanced text mining and semantic network analysis techniques to assess the impact of economic keywords on consumers' opinions and expectations regarding the economic situation and the Consumer Confidence Index (CCI) in Italy. This study used BERT to process and encode news, transforming large textual datasets into understandable formats that advanced predictive models can analyze. In particular, BERT allows the extraction of complex embeddings (vector representations) that capture the frequency of words and their semantic relevance and relationships with other words in the text. The embeddings generated by BERT were used to construct a word network, which was then analyzed using social network analysis methods to identify keywords' centrality and semantic importance in the economic field. This approach made it possible to quantify the impact of specific words on consumers' perceptions of the economic climate. The 'semantic importance' indicator used, Semantic Brand Score (SBS), combined word prevalence, diversity, and connectivity measures to provide a comprehensive metric beyond simple sentiment analysis. Integrating BERT and advanced embedding analysis has opened new frontiers in predicting and understanding consumer trust. These methods provide a valuable alternative to traditional surveys, often limited by response bias and delays in data collection. Using BERT and semantic analysis, it is possible to obtain a near real-time measure of the climate of trust, with potential applications ranging from economic crisis prevention to marketing strategies based on more accurate and timely macroeconomic data. The work effectively demonstrates how advanced NLP techniques, such as those offered by BERT, can be instrumental in analyzing and predicting economic behavior in an innovative and highly detailed manner, thus representing an essential benchmark for future studies in text-based economic prediction.

3. Methodology

The methodologies within the Word Vector Spaces ecosystem can broadly be categorized into two distinct families:

[1] More details are available at the following link: https://www.istat.it/en/experimental-statistic/social-mood-on-economy-index-2/.

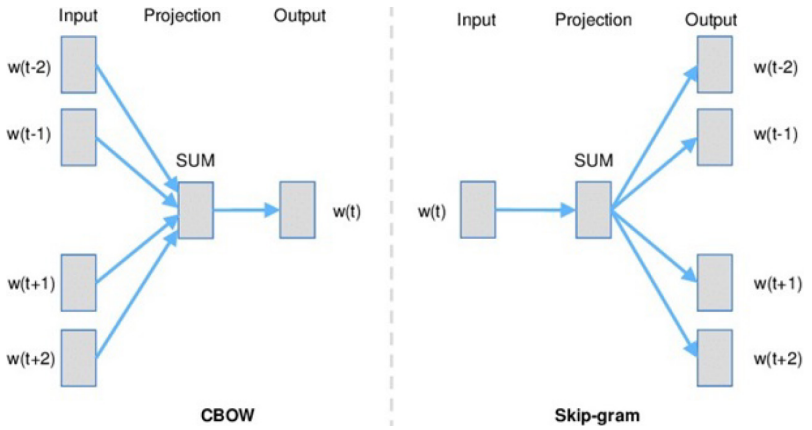


Fig. 1. Word2Vec architecture.

– Word Embeddings: unsupervised models that generate a vector space where each word is represented by a unique embedding that combines the word's different meanings into a single vector. Notable methods in this category include Word2Vec, GloVe, and FastText. Word2Vec, developed in 2013 by Google's team, is a toolkit designed to train vector space models for generating word embeddings [6]. Global Vectors for Word Representation (GloVe) learns context and word vectors by factorizing a global word-word co-occurrence matrix [7]. FastText, introduced in 2017 by Facebook's AI research team, enables the training of models on large corpora and computes word representations even for words not present in the training data [8].
– Context-Aware Word Embedding Methods: these can produce different embeddings for the same word, capturing the word's context, i.e., its position within a sentence. A prime example is the Bidirectional Encoder Representations from Transformers (BERT) [9].

Another prominent unsupervised learning task is Probabilistic Topic Modeling, utilized to extract latent semantic structures, often related to topics, from extensive text bodies. The most popular methods include Probabilistic Latent Semantic Analysis (PLSA) [10] and Latent Dirichlet Allocation (LDA) [11]. While word embeddings are prediction-based models, these topic modeling methods are count-based, identifying similar terms through their frequency across different documents. Similarities can be quantified using various similarity metrics in word embeddings and PLSA/LDA approaches.

3.1. Word embeddings

The most popular algorithm for learning word embeddings is Word2Vec; it harnesses the power of training a "shallow" neural network to learn word vectors. Word2Vec can alternatively adopt two different algorithms: Skip-gram, which predicts the context given its central word, and Continuous Bag of Words (CBOW), which predicts the central word given its context, as shown in Fig. 1.

In both Skip-gram and CBOW, words are one-hot encoded, and at the end of the training process, what is extracted is not the predictive output but its internal structure of weights.

Word2Vec offers several hyper-parameters for fine-tuning to enhance the quality of the learned model. These hyper-parameters are crucial in understanding Word2Vec's advantages over classical language modeling techniques such as Term-Document Matrix (TDM) or Term frequency – Inverse document frequency (Tf-Idf) [12]. Additional factors influencing Word2Vec model performance include size and quality of the corpus and the domain specificity: generally, larger corpora with less noise and fragmentation and with a specific domain context yield better performance and higher quality embedding spaces. At the end of the training phase, WEs appear as clusters of n-dimensional vectors showing words such that semantically similar words are closer within the vector space. These words' proximity is usually measured by space metrics such as the cosine distance, which quantifies similarity based on the cosine of the angle between two vectors.

Another algorithm to generate WE models is Global Vectors (GloVe); it merges the benefits of two primary natural language processing model families: global ma-

trix factorization (like latent semantic analysis, LSA) and local context window methods (like Word2Vec). Global matrix factorization models, which rely on term-document matrices, excel at utilizing statistical text data (such as word counts and frequency distributions), but fall short in solving analogies. In contrast, local context window models like Word2Vec aim to solve analogies but, before GloVe, did not explicitly use statistical information. GloVe combines these approaches by considering both the local context window of Word2Vec (where words occurring in similar contexts are syntactically/semantically related) and global-level co-occurrence statistics in the corpus. Unlike Word2Vec, GloVe explicitly incorporates these co-occurrence statistics into its algorithm, creating a model that leverages local context and global word-word co-occurrence.

In 2016, researchers at Facebook developed FastText, an extension of the Word2Vec model designed to overcome some of the limitations of traditional word embedding models such as Word2Vec and GloVe. Unlike its predecessors, which generate a single embedding for each word, FastText considers subword structures within words, i.e., word fragments or n-grams of characters. This allows FastText to generate vector representations for the words not seen during training (out-of-vocabulary), breaking them down into their n-character embeddings and aggregating them to form the final vector [13].

3.2. Top2vec

Top2Vec is one of the most important methods for Topic Modelling, a popular unsupervised learning task used to extract latent semantic structures, usually related to topics, from huge corpora of documents. While the traditional topic modeling algorithms such as Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) are native Bayesian, recently, topic modeling approaches that use WE spaces and then build clusters within the word embedding spaces have been introduced.

In 2020, another interesting embedding space was introduced with the name of Top2Vec, leveraging joint document and word semantic embedding to find topic vectors. Top2Vec is based on Doc2Vec, which extends the Word2Vec model to learn document-level representations. In a broad sense, Top2vec is an extension of Word2Vec, which leverages document and word embeddings to estimate distributed representations of topics. Top2vec takes document embeddings learned from a doc2vec model and reduces them into a lower dimension using an algorithm such as Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [14]. It then clusters the document vectors through algorithms such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [15]. This model does not require stop-word lists, stemming, or lemmatization, and it automatically finds the number of topics, unlike LDA, where this parameter must be manually set. The resulting topic vectors are embedded with the document and word vectors, and the distance between them represents semantic similarity. Top2vec finds topics significantly more informative and representative of the training corpus than probabilistic generative models. One major drawback of this methodology is that the number of clusters tends to be extremely high, requiring a hierarchical reduction.

3.3. BERTopic

BERTopic is an advanced topic modeling method that exploits text representations generated by transformer-based language models, such as BERT (Bidirectional Encoder Representations from Transformers), to identify and organize topics within large text corpora. In contrast to traditional topic modeling methods based on counting approaches such as LDA, BERTopic employs semantic embeddings to capture subtle linguistic nuances, thus enabling a more granular and contextually relevant classification of topics.

BERTopic method employs dimensionality reduction techniques, such as UMAP, and clustering algorithms, such as HDBSCAN, to identify clusters of documents that share similar thematic content. This approach offers a more precise and differentiated view of the topics emerging from the data. Further, it enhances the quality of thematic clustering and enables analysts to examine the evolution of topical dynamics over time and identify emerging trends in the analyzed corpus [16]. BERTopic differs significantly from Top2Vec in terms of approach and underlying technology. The results obtained with BERTopic are generally more interpretable than those produced by models based on traditional word embeddings, as BERT provides richer and more contextualized representations.

However, Top2Vec may be faster and more scalable, especially when using less complex representations such as Word2Vec. Due to its advanced contextual capabilities, BERTopic is often preferred over Top2Vec for analyzing complex texts.

Name	Weight
conflitto	0.6013
ucraina	0.6002
crisi	0.5942
putin	0.4990
europa	0.4771
recessione	0.4631
russia	0.4488
economia	0.4400
sanzioni	0.4371
guerre	0.4353

Fig. 2. Graph analysis and affinity list for word GUERRA (RUC model).

4. Results

The objective of this section is to describe the analysis results using both wordembox and topic modeling methods. In Section 4.1, we manually investigate the WEs of some keywords, their graphs, and how their representation changes over time. We also study how word embedding models trained on different datasets differ over time. As each word has a unique representation in every WE model and therefore provides a qualitative assessment of implicit clusters, in Section 4.2 we compare a Top2Vec analysis on WE spaces with classical topic modeling approaches such as Biterm. In Section 4.3, we perform topic analysis using BERTopic.

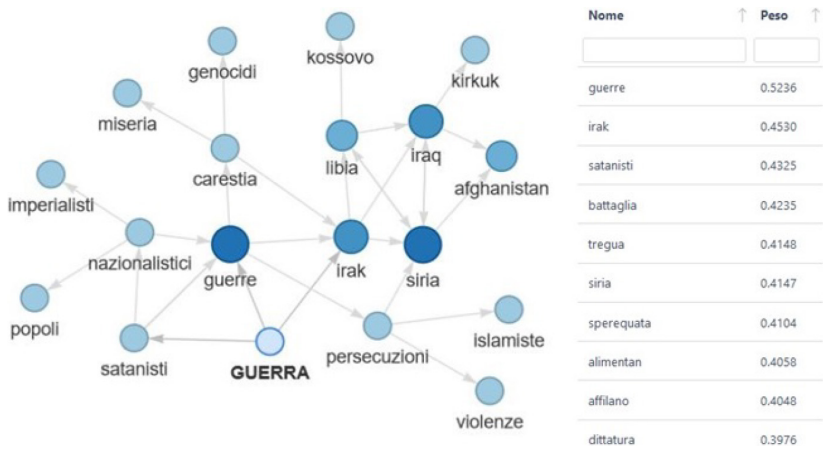
We aim to evaluate whether we can depict coherent topics with the two methods. The case study's purpose is to understand the Ukrainian crisis impact on social mood in economy posts collected using the Istat economic filter.[2] Therefore, we choose the posts collected from 20th February 2022 (a few days before the start of the Russia-Ukraine conflict) to 20th March 2022, consisting of 855.865 posts, as the reference dataset, called RUC. We note that the filter was not designed to sample posts related to the war, and only 10% of the sample posts contain the word *guerra* (war). In the first sub-section, we compare the word embedding model trained on the dataset with a model trained with posts from June 2016 to June 2018 (in the following SME17). For both trainings, we used a CBOW model (*embedding space dimension* $= 200$, *window size* $= 8$, *iterations* $= 20$). We focus on the words: *war*, *prices*, and *banks*. By analyzing the WE representation in February – March 2022, we observed the possible existence of several war-related topics. Therefore, we decided to perform further analysis by splitting the reference dataset into two periods of equal length. We performed a traditional LDA topic modeling approach (Biterm) because the WE models do not allow quantitative representations. We compared it with Top2Vec, which performs clustering in a word embedding space conceptually similar to Word2Vec. Our analysis aims to show how these topic modeling techniques perform compared to WE.

4.1. Russia-Ukraine conflict: Word embedding analysis

In this case study, we use the functionalities of wordembox to explore Italian X users' discussion about the Russian-Ukraine conflict and their concerns about the economy. Figure 2 shows the graph analysis and the list of affinity words of the word *Guerra* (War). The graph shows two areas: the first area (in the bottom of figure) concerns the conflict and geopolitical aspects, i.e., the words *conflitto* (conflict, also connected with other areas), *Ucraina*, *Europa*, *Russia*, and *Putin* (Ukraine, Europe, Russia, Putin); the second area (on the right of figure) concerns the consequences of the war on the Italian economy, i.e., *crisi*, *recessione* and *inflazione* (crisis, recession, and inflation).

We obtain very different conclusions analyzing the same graph in model SME17 (Fig. 3). Also, in this case, the figure shows two areas: at the top right, the geo-political area where we find the words *Siria*, *Iraq*, *Afghanistan*, *Libia* and so on; the area in the left and in the bottom displays consequences and causes of the conflicts i.e., *nazionalistici*, *carestia*, *persecuzioni*, *vio-*

[2]More detailed information about the Italian Social Mood on Economy Index (SMEI) and the filter, i.e. a list of relevant words, can be found at https://www.istat.it/wp-content/uploads/2018/07/methodological-note-social-mood.pdf.



Nome	Peso
guerre	0.5236
irak	0.4530
satanisti	0.4325
battaglia	0.4235
tregua	0.4148
siria	0.4147
sperequata	0.4104
alimentan	0.4058
affilano	0.4048
dittatura	0.3976

Fig. 3. Graph analysis and affinity list for word GUERRA (SME17 model).

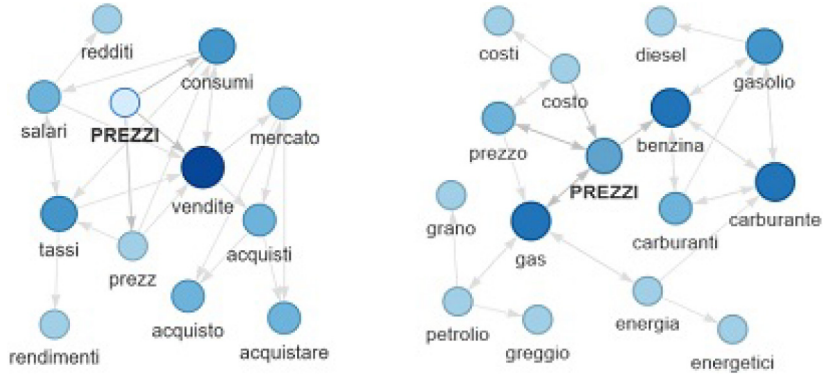


Fig. 4. Graph analysis for the word PREZZI: SME17 model (left), RUC model (right).

lenze and *genocidi* (nationalities, famine, persecutions, violence and genocides).

In addition to the geopolitical differences that characterize the war perceived by Italian X users, it is evident how the related topics range from humanitarian aspects to economic aspects. Further, we compare the graph analysis of the words *prezzi* (prices) and *banche* (banks) in the two models. Figure 4 concerns prices. In the RUC model (graph on the right), X users express their concerns related to the increase in costs of fuel and gas; indeed, we find the words gas, *benzina* (gasoline) and *carburante* (fuel). The conversations in SME17 model (graph on the left) concentrated on the purchasing power of salaries in buying and goods consumption; indeed, we find the words *vendite* (sales), *consumi* (consumptions), *tassi* (rates) and *salari* (wages).

Concerning banks, in Fig. 5 we observe that while in RUC the debate on banks focuses on sanctions for Russian account holders, in the SME17, the focus was on the crisis witnessed by some important Italian banks: *Monte dei Paschi* (MPS) and *Banche Venete*. Indeed, in the graph on the right (RUC), we find the words *Russia*, *swift* (word network banks system) and *sanzioni* (sanctions); instead, in the graph on the left (SME17) we find the words *risparmiatori* (savers), *correntisti* (account holders), *mps* and *venete* (the two Italian banks).

4.2. Russia ukraine conflict: Topic modeling

In the previous section, we analyzed word embedding graphs and affinities for a one-month period. In this section, we characterize groups of words related to two contiguous periods by splitting the first 4 weeks of the war in Ukraine into two periods: 20th February – 6th March (P1) and 7th March – 20th March (P2). We aim to analyze the dynamics of the terms from P1 to P2 and evaluate whether different topic modeling techniques (Biterm and Top2Vec) can produce coherent representations. For comparability, we set the same number of topics, namely 10, in the two topic modeling approaches for both periods.

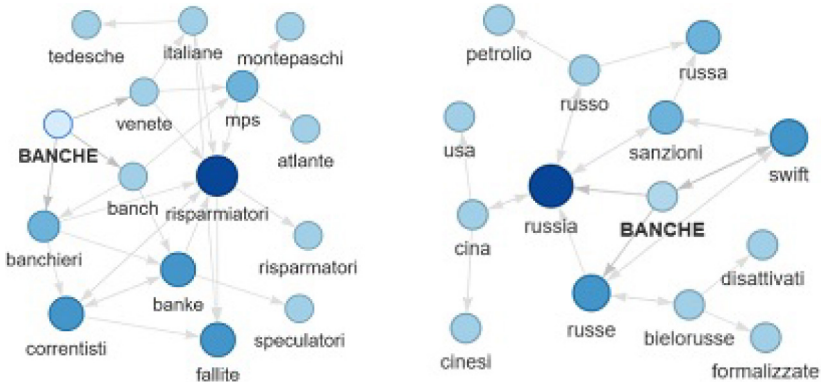


Fig. 5. Graph analysis for the word BANCHE: SME17 model (left), RUC model (right).

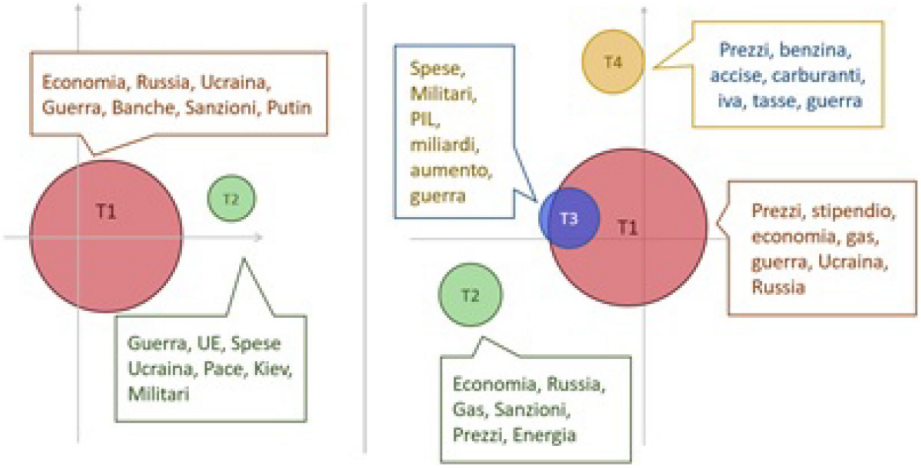


Fig. 6. Biterm clusters containing the term war: P1 (left) and P2 (right). P1 topics in English: T1 = Economy, Russia, Ukraine, War, Banks, Sanctions, Putin; T2 = War, EU, Ukraine Expenditure, Peace, Kiev, Military. P2 topics in English: T1 = Prices, salary, economy, gas, war, Ukraine, Russia; T2 = Economy, Russia, gas, sanctions, prices, energy; T3 = Expenditure, Military, day, GDP, billions, increase, war; T4 = Prices, petrol, excise duties, fuel, VAT, taxes, war.

First, we will comment on the results obtained using Biterm. In this case, we selected the major clusters containing the word *war*, *Russia* to reach 80% of the total occurrences of the word war and Russia in both P1 and P2. This topic modeling, shown in Fig. 6, displays the word clusters as circles whose diameter is proportional to the size of the cluster, where the total size is the count of occurrences of words in the document term matrix. Regarding the absolute frequency of the term war, it can be observed that it is almost constant in the two periods, with the number of occurrences slightly above 40,000 in the first period and slightly less in the second. We observe that more topics are related to the conflict in the second period. Indeed, there are two topics in the first period and four in the second. In particular, in P1, the clusters T1 and T2 cover 63.9% of the total word occurrences of the corpus, while in P2, the topics T1, T2, T3, and T4 cover 67.1% of the total word occurrences of the corpus. In the first period, the predominant topic contains the following keywords: *economy*, *Russia*, *Ukraine*, *banks*, and *sanctions*.

If we analyze the topics in period P2, we observe that the main topic, T1 (46%), is related to prices and gas. The topic on military expenses is observed in both periods, P1 (T2 5%) and P2 (T3 6.7%). In the first phase, economic opinion (T1 59%) focuses almost exclusively on the problem linked to sanctions, Russia, and banks. In contrast, in the second period, conversations focus on the problem linked to the increase in gas, energy, and fuel prices (T1 46%, T2 7.5%, T4 7%) for a total of 60.5% of posts.

We conducted a further analysis using Top2Vec, employing its hierarchical reduction technique to reduce the number of topics to 10 in both periods (T1 and T2).

Fig. 7. Word cloud depicting P1 clusters: 6.8% cluster related to banks and sanctions (top); 13.7% cluster related to inflation and gas (bottom).

Fig. 8. Word cloud depicting clusters related to military expenses and weapons: 6.8% P1 cluster (top); 20.6% P2 cluster (bottom).

It is important to note that initially, Top2Vec identified over 1400 topics, which required the use of hierarchical reduction. Within this approach, the sizes of clusters are more balanced when compared to Biterm, ranging from 22.3% to 6.8% in P1 and from 20.6% to 5.2% in P2. In this case, we obtained that in period P1, there are 6 topics related to war and its economic consequences covering 53% of the posts. We can recognize two clusters related to banks (10%) and to sanctions (6.8%), and in this case, a cluster related to economy, inflation and gas (13.7%), that with Biterm doesn't explicitly appear in the first period P1, and one about military expenses 6.8% (see Fig. 7).

In the second period, we observe six clusters related to war covering 62.5% of posts. More precisely, we find a 20.6% cluster on military expenses (see Fig. 8) and two clusters related to the increase of prices in fuel (13.6%) and energy in general (8.2%), respectively, top and bottom in Fig. 9.

4.3. Russia-Ukraine conflict: BERTopic analysis

To maintain comparability with the analyses of the previous paragraphs, we conducted a BERTopic analysis on the same parameters. We, therefore, considered the same periods, P1 and P2, and set a maximum number of topics to 10, after which we focused only on the first two topics that were most relevant to the topic *war*. In both periods, the topic *war* was predominant. The first two topics, in terms of frequency of occurrence, concern the topic *war* and account for approxi-

Fig. 9. Word cloud depicting T2 clusters: 13.6% cluster related to energy (top); 8.2% cluster related to fuel price increase (bottom).

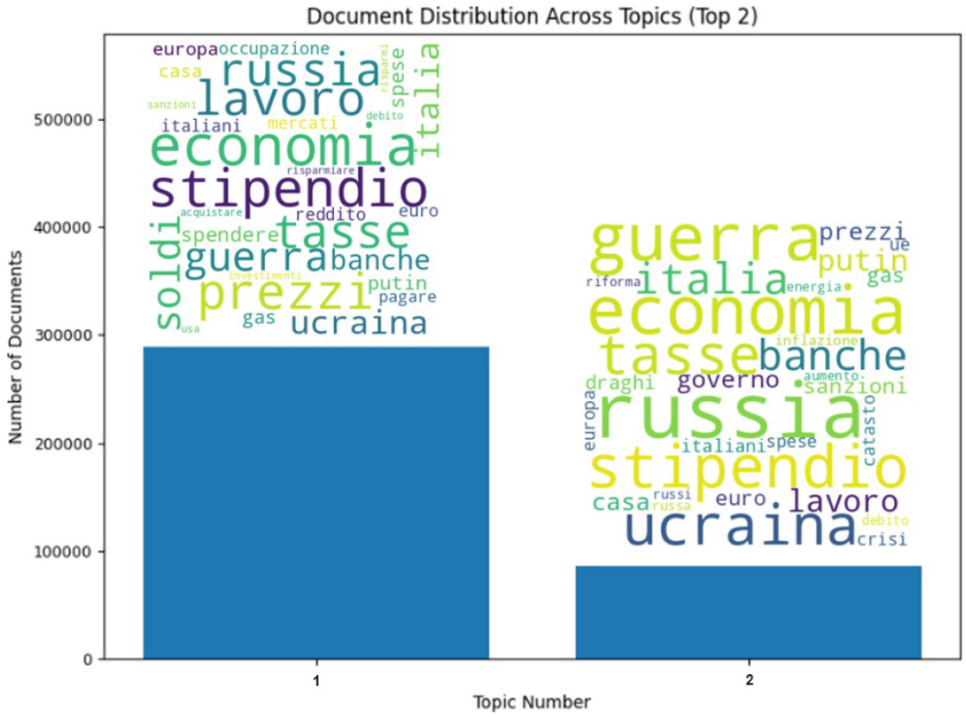


Fig. 10. Histogram and word cloud representation of topic 1 and topic 2 related to the war in period P1.

mately 98% of the total number of posts (equivalent to BERTopic documents) in the corpus. In absolute terms, in period P2, the first two topics increased by 17.9%.

The word clouds in Figs 10 and 11 demonstrate no substantial differences between the two main classes, namely the first two topics in the first and second period. When the two periods are considered jointly, it becomes evident that the relevant words in the period P1 are: *Russia*, *economy*, *war*, *salary*, *Ukraine*, *taxes*, *prices*, *work* and in the period P2 are: *prices*, *economy*, *Italy*, *Russia*, *war*, *Draghi*, *taxes*, *salary*, and *work*. The topics' content show that the discussion on the war and its economic impacts was dominant in both periods. However, in period P2, there is a greater emphasis on prices and economic conditions in Italy. Furthermore, the number of documents (posts) associated with the main topics increased in the second period, indicating a growing concern and discussion on these issues. The

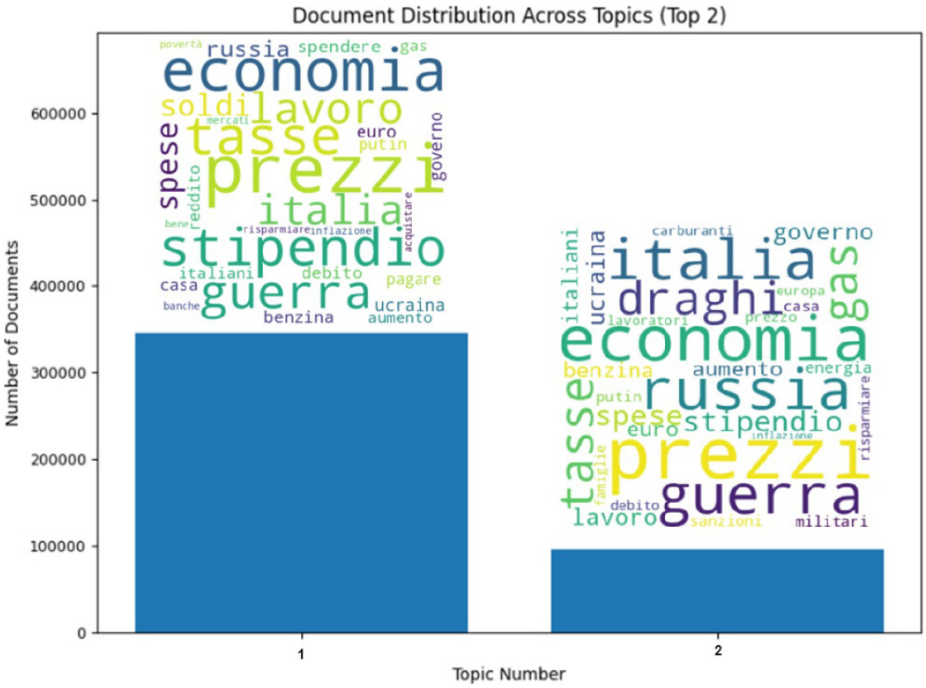


Fig. 11. Histogram and word cloud representation of topic 1 and topic 2 related to the war in period P2.

analysis suggests that while the main topics remain consistent, the intensity and specificity of the discussions have evolved, with a greater emphasis on direct economic impacts, such as prices and government policies.

5. Conclusion

Integrating word embedding and topic modeling techniques for sentiment analysis offers a powerful approach to understanding social mood, especially during significant events like the Russia-Ukraine conflict. This study highlights the effectiveness of combining traditional and advanced natural language processing (NLP) methods to analyze social media data, specifically X posts, to assess public sentiment and topic prevalence.

Utilizing Istat's wordembox, this research demonstrated the capability to visualize and interpret the semantic relationships between words within embedding spaces. This visualization is crucial for identifying implicit clusters and understanding the context and evolution of key terms over time. The comparison between word embedding models trained on different datasets revealed significant shifts in topic associations, particularly related to the economic impact of the conflict. Moreover, the study compared the performance of various topic modeling techniques, including Bayesian Latent Dirichlet Allocation (LDA), Top2Vec, and BERTopic. Each method provided coherent and contextually relevant results, underscoring the potential synergy between these techniques in enhancing the informative content of synthetic indexes like the Social Mood on Economy Index (SMEI).

The detailed analysis of posts from the onset of the Russia-Ukraine conflict showed a distinct shift in public discourse from general geopolitical concerns to specific economic anxieties, such as rising fuel and gas prices. This shift was effectively captured through traditional topic modeling methods and the more nuanced approaches of Top2Vec and BERTopic. The latter's ability to utilize transformer-based embeddings allowed for a more granular and contextually rich classification of topics, offering deeper insights into the evolving public sentiment. In conclusion, the combined use of word embedding and topic modeling represents a robust framework for sentiment analysis, particularly in the context of dynamic and multifaceted social issues.

There are several reasons why these techniques are valuable and necessary when dealing with experimental statistics based on social media. One of the main critiques is that synthetic indexes cannot show or summarize the main topics discussed. Users' readability and official statistics' credibility may be increased using topic modeling techniques. In a typical social media

pipeline, selecting the right keywords is crucial for sampling texts containing at least one of these expressions. Analysis based on word embedding spaces is necessary to validate this choice through a data-driven bottom-up approach, enhancing the index's quality, i.e., relevance.

In conclusion, the combined use of word embedding and topic modeling represents a robust framework for sentiment analysis, particularly in the context of dynamic and multifaceted social issues. Future enhancements will involve integrating new and diverse methods to refine the analysis and expand its applicability across different domains. This study lays the groundwork for utilizing advanced NLP techniques to provide near real-time insights into public sentiment, thereby supporting more informed decision-making processes in both research and practical applications.

Official statistics are increasingly interested in investigating the possible use of new sources, including social media platforms, to enrich production. However, to fully exploit the potential offered by these new data sources, National Statistical Institutes must invest in methodological tools useful for the processing and quality evaluation of these data sources. Currently, Istat is applying topic modeling techniques to study opinions on immigration through X. Another experimental project is about online gender-based violence and hate speech.

Annex 1: Wordembox

As widely discussed in the literature [6], word embedding algorithms transform words into vectors of a low-dimensional metric space, usually set to values between 100 and 300. The word embedding model can contain hundreds of thousands of vectors for a large input corpus. As a result, the complete structure of the embedding model is very difficult to analyze. To explore and visualize such patterns, we need to (i) reduce the dimensionality of the embedding space and (ii) focus only on a subset of vectors, i.e., those derived from the words most relevant to the desired analysis.

The first task can be solved by traditional solutions for dimensionality reduction, such as Principal Component Analysis (PCA) or a stochastic variation such as t-SNE [7]. For the second, however, no standard methods are available. The general idea behind implementing the wordembox is to adopt a new technique based on graphs [6], simultaneously addressing both needs. The tool selects only a subset of relevant words by adopting a filtering criterion based on their semantic proximity while at the same time allowing the visualization of the resulting sub-model in a two-dimensional graph to be easily represented and readable.

Wordembox is a tool implemented by Istat [8] designed to delve into WE models. The tool facilitates such tasks by leveraging the model's inherent capabilities, such as affinity and analogy tests, and visualizing the connections between words via graphs. The draft version, documented in [16], encompassed the features. A detailed description of the functionalities provided by wordembox is provided in below. However, the original version of wordembox was affected by several limitations: (i) exclusive accessibility to Istat's internal researchers, (ii) its source code remained inaccessible to the public, (iii) the ability to handle only a single WE model at a time, and (iv) the reliance on outdated technological frameworks.

The latest version of wordembox, has successfully solved these challenges. This updated release supports the simultaneous management of various WE models and empowers users with the flexibility to select and examine any model of their choosing. Moreover, the source code is publicly accessible on GitHub.[3] The architectural diagram, shown in Fig. 12, outlines Istat's multi-stage process for handling X data. Below is a breakdown of wordembox's components and data processing workflow.

- X Data: the process starts with data retrieved from the social platform X. Data are downloaded from the platform using a filter, i.e., a set of keywords aimed at extracting posts that are of interest for the analysis.
- Data Pre-processing (Batch): this is the first major phase in the workflow where the X data undergoes several steps: 1) *text pre-processing*: convert text to lowercase, tokenize the text into words, apply basic orthographic repairs, remove URLs, remove non-alphabetic characters (e.g. '#' or '@'), remove stop words, if needed, stem words to get rid of inflected forms. 2) *Word Embedding model*: the post corpus is fed into a word embedding model. During this step, the model parameters are optimized to convert words from the posts into vector representations.
- Data Analysis (Backend): The data moves to the analysis phase once the word embeddings are generated. The backend provides a set of function-

[3] The source code of wordembox is available at the following link: https://github.com/istat-methodology/nlp.

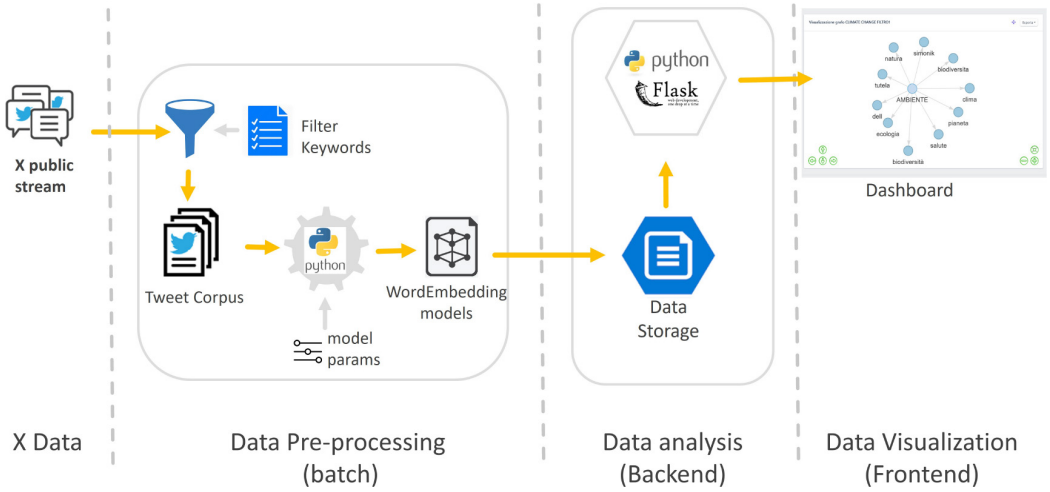


Fig. 12. Wordembox architectural components.

alities that allow the analysis of the WE models. A detailed description of such functionalities will be provided in the following sections. The data storage component is a repository for the raw post corpus, the generated word embeddings, and any intermediate results from the analysis phase.

– Data Visualization (Frontend): Finally, the analysis results are visualized in a dashboard. The visualization includes network graphs to show relationships or associations between words based on the word embeddings. This dashboard is the tool's frontend, allowing users to interact with the models and interpret the analysis results.

Affinity functionality

This functionality, initiated by one or several seed words, generates a list of words closest in syntax or semantics by measuring their proximity in the embedding space. The function receives in input two parameters:

1. Seed Word(s): This parameter specifies the word(s) or vector(s) relative to which the algorithm searches for the n closest words or vectors within the embedding space, using the cosine distance as a measure. Additionally, it allows for the input of multiple words. When multiple seed words are provided, the search for associated words is based on the aggregate vector of the inserted words or vectors. This is particularly useful for addressing disambiguation issues, where a word may have multiple meanings. For example, *Rome* could relate to semantic contexts such as historical, geographical, or sports-related references.
2. Number of words: This parameter defines the quantity of related words the user wants to retrieve.

Analogy functionality

Solving analogies represents one of the most impressive capabilities of word embedding models, hinging on vector arithmetic to discern semantic relationships as vector differences between pairs of words. For instance, if provided with two words that share a specific relationship and a third word to extend this relationship to (e.g., *man* is to *king* as *woman* is to *x*), the model can generate a list of words that aptly complete the analogy (with *x* being *queen*). This feature is made possible by the word embedding model's ability to represent words as vectors.

To utilize the analogy function, the following parameters are required:

– Number of words: This parameter determines how many words the model should offer as potential solutions for completing the analogy.
– word1, word2, word3: These are the three words that form the basis of the analogy.

Graph functionality

This functionality enables the visualization of the WE model on a two-dimensional canvas, using one or more seed words as the starting point to define the semantic territory for exploration.

The expressive power of graphs has been integrated into the wordembox dashboard. We've developed three

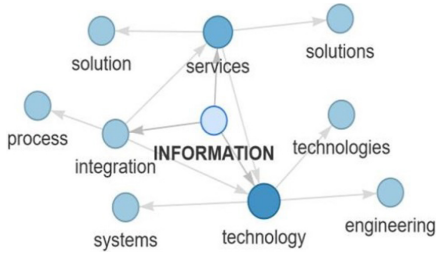


Fig. 13. Example of graph analysis using the seed-word INFORMATION.

distinct methods to assemble basic graphs tailored to different exploration strategies: geometric, linear, and geometric-oriented. Each graph type calls for specific input parameters: width, number of iterations, and one or more seed words. The *width* parameter controls the number of words or nodes close to the seed word – or, in subsequent iterations, close to the newly discovered words – to be included in the graph at each stage. The *iterations* parameter sets how many cycles of exploration are desired, while the *word* parameter specifies the initial seed word(s) for beginning the exploration. Additionally, wordembox incorporates two further parameters: *mode* and *layout*, which dictate the visual presentation and the structural type of the graph (geometric, linear, or geometric-oriented), respectively.

- Geometric graph: starting from a *seed word*, at each iteration, the following are displayed: in the first iteration, the *n* words closest to the *seed word*, and in subsequent iterations, the *n* words closest to the words found in the previous iteration; so, the *geometric graph* tends to expand the range of exploration very quickly, rapidly losing the initial semantic focus provided by the *seed word*.
- Linear graph: each iteration introduces a virtual node that represents the aggregate of the previously discovered words. The number of nodes increases linearly with each iteration. This graph is straightforward to interpret and tends to chart a *semantic trajectory* through the embedding space, akin to narrating a story.
- Geometric oriented graph: this often serves as an optimal balance between the former two types. Here, every discovered word at each iteration generates a virtual node, aggregating all words from non-virtual nodes found along the shortest path, linking the current word back to the seed word.

Figure 13 displays a semantic graph generated using wordembox, which visualizes relationships between the central concept INFORMATION and related terms. The central node, INFORMATION, is the largest and is highlighted, indicating its primary importance or focus within the graph. Surrounding it are nodes representing related concepts such as *services*, *solutions*, *technologies*, *integration*, *systems*, *process*, *solution*, and *engineering*. Each node is connected to INFORMATION via edges, suggesting a direct relationship between each term and the central concept. Some nodes are interconnected, indicating a semantic or syntactic relationship between these terms. For example, *solution* and *solutions* are likely connected because they are grammatical variations of the same concept, and their proximity to INFORMATION suggests they are relevant in the context of this WE model. Using varying node sizes implies a difference in relevance or frequency; larger nodes, like *technology*, represent more prominent or frequent associations with INFORMATION within the dataset. The graph structure allows for an intuitive understanding of how different concepts are interrelated around the central theme of INFORMATION, which can be particularly useful for exploring semantic fields or discovering new associations within a corpus of text.

References

[1] Catanese E, Bruno M, Scannapieco M, Valentino L. Natural language processing in official statistics: The social mood on economy index experience. Statistical Journal of the IAOS. 2022; 38(4): 1451-1459. doi: 10.3233/SJI-220062.

[2] Mikolov T, Yih W-T, Zweig G. Linguistic regularities in continuous space word representations, in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (Vanderwende L, Daumé III H and Kirchhoff K, eds.), (Atlanta, Georgia). 2013; 746-751. Association for Computational Linguistics.

[3] Firth HWJ. Papers in linguistics 1934-51. International Journal of Applied Linguistics. 2007; 402-413.

[4] Bruno M, Scannapieco M, Catanese E and Valentino L. Italian sentiment analysis on climate change: Emerging patterns from 2016 to today. Statistical Journal of the IAOS. 2023; 39(1): 189-202. doi: 10.3233/SJI-220064.

[5] Fronzetti Colladon A, Grippa F, Guardabascio B, Costante G, Ravazzolo F. Forecasting consumer confidence through semantic network analysis of online news. Scientific Reports. 2023; 13(1): 11785.

[6] Gibbons A. Algorithmic graph theory. Cambridge university press. 1985.

[7] Van der Maaten L, Hinton G. Visualizing data using t-sne. Journal of Machine Learning Research. 2008; 9(86): 2579-2605.

[8] De Fausti F, De Cubellis M, Zardetto D. Word embeddings: A powerful tool for innovative statistics at istat. in Proceedings of the 14th International Conference on Statistical Analysis of Textual Data – JADT. (Rome, National Research Council). 2018; 571-583.

[9] Bruno M, Catanese E, De Cubellis M, De Fausti F, Pugliese F, Scannapieco M and Valentino L. Analyzing textual data

through word embedding: experiences in istat. in Book of Short Papers, 51th Scientific Meeting of the Italian Statistical Society. 2022; 571-583. PEARSON.

[10] Hofmann T. Probabilistic latent semantic indexing. in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '99, (New York, NY, USA). 1999; 50-57. Association for Computing Machinery.

[11] Blei DM, Ng AY, Jordan MI. MILatent dirichlet allocation. J Mach Learn Res. 2003 Mar; 3: 993-1022.

[12] Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics. 2015; 3: 211-225.

[13] Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. CoRR. 2016; vol. abs/1607.01759.

[14] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. 2020.

[15] McInnes L, Healy J and Astels S. Hdbscan: Hierarchical density based clustering. The Journal of Open Source Software. 2017 Mar; 2: 205.

[16] Grootendorst M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794. 2022.