IJCoL
Italian Journal of Computational Linguistics

8-2 | 2022
Italian Journal of Computational Linguistics vol. 8, n. 2 december 2022

Word Usage Change and the Pandemic: A Computational Analysis of Short-Term Usage Change in the Italian Reddit Community

Edoardo Signoroni, Elisabetta Jezek and Rachele Sprugnoli

Electronic version
URL: https://journals.openedition.org/ijcol/1076
DOI: 10.4000/ijcol.1076
ISSN: 2499-4553

Publisher
Accademia University Press

Electronic reference
Edoardo Signoroni, Elisabetta Jezek and Rachele Sprugnoli, "Word Usage Change and the Pandemic: A Computational Analysis of Short-Term Usage Change in the Italian Reddit Community", *IJCoL* [Online], 8-2 | 2022, Online since 01 December 2022, connection on 23 February 2023. URL: http://journals.openedition.org/ijcol/1076 ; DOI: https://doi.org/10.4000/ijcol.1076

Word Usage Change and the Pandemic: A Computational Analysis of Short-Term Usage Change in the Italian Reddit Community

Edoardo Signoroni*
Università di Pavia, Masaryk University

Elisabetta Jezek**
Università di Pavia

Rachele Sprugnoli†
Università di Parma

The COVID-19 pandemic has affected every aspect of our lives. Our work assesses whether it has also impacted the usage of the Italian language, particularly its lexicon. We create a new corpus of Italian texts taken from Reddit and apply a recent unsupervised usage change detection method on two sub-corpora, one with data from 2019 and one with data from 2020. The focus of our investigation is short-term usage change. The results for the first 10-top candidates and for a selection of candidates among the top-100 are analyzed, to show that usage change has indeed happened.

1 Introduction

Each aspect of language changes over time, but meaning is the one more susceptible to mutation. According to Blank (1999), change is only a side-effect of the speakers' pragmatic goal, which is to achieve success in communication. This also means that change is a consequence of the human mind and social interactions: innovations are thus employed and adopted because they are judged to be the most successful strategy to communicate effectively.

The study of meaning change was the focus of the first scholars of semantics but while they employed manual methods, nowadays many studies are conducted with automatic and semi-automatic tools stemming from computational linguistics and computer science.

Lexical Semantic Change (LSC) detection, which aims at identifying the change in meaning of words over time using corpus data, is a Natural Language Processing (NLP) task pertaining to lexical and diachronic semantics. Recently, this field has seen an exponentially rising interest but work for languages other than English is still relatively scarce (Schlechtweg et al. 2020).

The computational literature approaches the task in several ways and with different terminologies: Tahmasebi, Borin, and Jatowt (2021) define the field as "lexical semantic change detection"; this definition is also adopted by both Schlechtweg et al. (2020) and

* NLP Centre, Faculty of Informatics - Botanická 68a, 602 00 Brno, Czech Republic. E-mail: e.signoroni@mail.muni.cz
** Dipartimento di Studi Umanistici, Corso Strada Nuova 65, 27100 Pavia, Italy. E-mail: jezek@unipv.it
† Dipartimento di Discipline Umanistiche, Sociali e delle Imprese Culturali, Via M. D'Azeglio, 85, 43125 Parma, Italy. E-mail: rachele.sprugnoli@unipr.it

Basile et al. (2020b), which set the task as "identifying words that change meaning over time". Kutuzov et al. (2018), instead, formalize the task as "detecting semantic shifts". Finally, Del Tredici, Fernández, and Boleda (2019) employ "short-term meaning shift", while Gonen et al. (2020) frame the task as "detecting usage change": in this paper we follow this latter definition.

Most of the work in this field studies meaning change across decades or even centuries, by leveraging data from different corpora of literary or newspaper data. Fewer studies investigate short-term usage change, by comparing texts produced in smaller time spans, from one to less than ten years apart. This kind of research often uses data from social media, like Twitter or Reddit. When considering such smaller time frames, it is more sensible to talk about "usage change" rather than "meaning change" of a word, as proposed by Gonen et al. (2020).

The focus on use is motivated by the distributional method adopted to investigate the data (Harris 1954), which derives information about the meaning of a word from its context of use, and assumes that words with similar distributional properties have similar meanings (Sahlgren 2008; Ježek 2016; Lenci 2018; Jurafsky and Martin 2021). The kind of semantics that stems from the distributional hypothesis is called distributional semantics or, more specifically, vector space semantics, because it represents words and their meaning as vectors in a geometric space, and calculates the similarity between vectors using the cosine function. With cosine similarity, the nearest neighbors, i.e. the items with the highest similarity score with respect to the target word, can be identified (Lenci 2018). In distributional semantics, there are several types of vectors that are computed with different methods, in particular count-based vectors obtained by counting the co-occurrences of words, and embedded vectors (called *embeddings*) obtained with predictive neural models. The vectors we use in our experiment belong to the second type and are computed with the *Skip-Gram with Negative Sampling* (SGNS) version of the *word2vec* neural model (Mikolov et al. 2013).

In our study we consider a short-term time span that represents a peculiar socio-cultural and chronological context, the pandemic. Our work starts from the hypothesis that an event such as the COVID-19 pandemic would bring forth changes in the use of words. We focus on Italian, since many other studies were done for the English language. To achieve our goal, we create a new corpus of texts from Reddit, and partition the corpus in two datasets, one for the year 2019 and one for the year 2020. After cleaning and lemmatizing the corpus, we apply the method outlined in Gonen et al. (2020) (§3.3) to our data to detect word candidates that may have undergone usage change from one dataset to the other.

Our analysis of the proposed candidates indicates that some degree of usage change has occurred: specific word senses gained prominence and new words arose as the need to express concepts connected to the pandemic became more widespread.

The paper is divided into five sections: Section 2 reviews the contribution of computational linguistics and Natural Language Processing (NLP) to the COVID-19 pandemic; it also formally defines the task of unsupervised meaning change detection and surveys different approaches. Section 3 details the methodology of this study and describes the features of our datasets. Section 4 presents the results of the work and analyzes them. Section 5 draws some conclusions.

Contributions

Our work contributes to the research on usage change detection and on the impact of the COVID-19 pandemic on language as follows:

- creating a new corpus of social media texts for Italian, focusing on short-term usage change. The corpus is available online;[1]
- testing the application of a relatively recent and computationally light method of usage change detection, previously untested for Italian;
- analyzing the impact of the pandemic on word use in a language different than English.

2 Related work

In this section we first provide an overview of the work done by the Computational Linguistics and NLP community in response to the COVID-19 pandemic (§2.1). Then, we briefly survey previous studies and methods of computational detection of meaning change (§2.2), with a focus on short-term change (§2.2.1) and on Italian (§2.2.2).

2.1 Computational Linguistics and the COVID-19 Pandemic

The Computational Linguistics and NLP community can support the research to fight the Coronavirus and its consequences by tapping into the great quantities of unstructured text and speech data; analyzing the countless published research papers, social media post and news articles can be critical to support best practices in clinical management; to understand the public response to the outbreak; to find and contrast spreading misinformation; to automatically identify and organize helpful information from the web.

One of the first resources on the COVID-19 pandemic is CORD-19, a COVID-19 Open Research Dataset[2] curated at the Allen Institute for AI in March 2020. In the same month, the 'Lab Task 1' at CLEF (Conference and Labs of the Evaluation Forum) 2020[3] asked to rank a stream of tweets on different topics, including COVID-19, according to their check-worthiness. A check-worthy tweet includes a claim that is of interest to a large audience or that might have a harmful effect. Again, in March, the Kaggle platform[4] started to organize tasks to develop text and data mining tools that can help the medical community to develop answers to high priority scientific questions. These are based on the aforementioned CORD-19 corpus, as is the TREC (Text Retrieval Conference)-COVID program[5], a challenge that follows the TREC assessment process to evaluate search systems.

In July 2020, the 1st Workshop on Natural Language Processing for COVID-19 was held at the Association for Computational Linguistics (ACL) conference. The second part of the workshop was held in the same year at the Empirical Methods in Natural Language Processing Conference (EMNLP). Both workshops demonstrated the help that the NLP community can provide, mainly in navigating the literature on the virus, in identifying and fighting misinformation and in characterizing the public reaction through the analysis of data from social media, like Twitter and Reddit.

One of the first contributions of the Italian NLP community to fight the pandemic is 40twita, part of the larger TWITA project ongoing at the University of Turin since

1 https://github.com/edoardosignoroni/usage_change_ITA
2 https://www.semanticscholar.org/cord19/download
3 https://clef2020.clef-initiative.eu/
4 https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
5 https://ir.nist.gov/covidSubmit/

2012. TWITA is a collection of tweets in Italian, first published in 2013 with about 100 million tweets from February 2012 to February 2013; the automatic collection is still ongoing. 40twita is a subset of TWITA; the dataset is collected daily from 1 March 2020 by filtering TWITA with COVID-19 related keywords.[6]

In April 2020, the "Covid-19 Semantic Browser" was developed by the Area Science Park in collaboration with the Italian Association of Computational Linguistics (AILC): it employs state-of-the-art neural networks to search relevant articles in the CORD-19 dataset.[7] Another useful tool developed by the Italian community is the "COVID19 Infodemics Observatory"[8] at the Complex Multilayer Networks (CoMuNe) Lab of the Fondazione Bruno Kessler in collaboration with Harvard's Berkman Center for Internet & Society and with IULM University in Milan. According to the WHO,[9] the pandemic has been accompanied by a massive surge of information, dubbed "infodemic", which can potentially contain inaccurate, fake, or harmful information. This makes it hard for people to find reliable and trustworthy news and sources. FBK's Observatory monitors millions of tweets with machine learning techniques to quantify collective sentiment and psychology, presence of social bots[10] and news reliability to find that almost 30% of the news are unreliable.[11]

2.2 Automatic Language Change Detection

Formally, the task of detecting meaning change can be formulated as follows: given corpora [C_1, C_2, ..., C_n] containing texts created in time periods [1, 2, ..., n], the task is to locate the same words with different meaning in different time periods, or to locate the words which changed the most. Related tasks are to discover general trends in meaning change or the dynamics of the relationships between words (Kutuzov et al. 2018).

At the word level, most of change detection methods employ vectors, both count-based and neural ones (embeddings), for the words. This comes to the cost of representing all senses of a term with a single representation. Most of the count-based approaches start by building a co-occurrence matrix, often reducing its dimensions by SVD (Singular Value Decomposition). PMI (Pointwise Mutual Information) scores are used for co-occurrence strength rather than raw frequency, while vector similarity is measured with the cosine (Tahmasebi, Borin, and Jatowt 2021). Low similarity is understood as higher amount of change or polysemy.

Sagi, Kaufmann, and Clark (2009) employ context vectors, that is, the combined vectors of the words in a context window around the word under examination, while Gulordava and Baroni (2011), and Rodda, Senaldi, and Lenci (2017) use also PMI. Kahmann, Niekler, and Heyer (2017) compare changes in context similarity between ranked series at different points in time. Tang, Qu, and Chen (2013) and Tang, Qu, and Chen (2016) use contextual entropy and reduce dimensions on the fly rather than through SVD. Most of these methods are evaluated qualitatively on a random or manually selected sample.

6 http://twita.di.unito.it/dataset/40wita
7 http://covidbrowser.areasciencepark.it/
8 https://covid19obs.fbk.eu/#/
9 https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports, Situation Report 13, 2 Feb 2020.
10 A social bot is an automated computer program that interacts with users on social media.
11 https://covid19obs.fbk.eu/#/

The works that use word embeddings train them independently over different time-sliced corpora and then compare them by projecting all representations onto the same space. More specifically, there are three main methods: i. vectors for the first time period are trained without any other information, then the representation for the successive time spans is initialized with the values of the previous interval to which they are then compared using cosine similarity to detect the change (Kim et al. 2014); ii. words are projected using linear mapping on the last time period (Kulkarni et al. 2014; Hamilton, Leskovec, and Jurafsky 2016); iii. mapping is avoided all together by comparing second order similarity, and meaning is modelled as the linear combination of the neighbors of a word from previous time points (Eger and Mehler 2016).

Dynamic word embeddings are another embedding method for meaning change detection. While different techniques exist involving these vectors, all of them train this kind of word embeddings in the same original space, and then share data across all time periods to update the word representations. Dynamic word embeddings have been shown to be beneficial, because they reduce the need of aligning independently trained embeddings, and the necessity of large datasets, rarely available for historical corpora (Tahmasebi, Borin, and Jatowt 2021). This approach is employed by Bamler and Mandt (2017), Yao et al. (2018), and Rudolph and Blei (2018). Context vectors are shared across all the time slices, while the embeddings are trained only within a single time span. It was shown that dynamic word embeddings perform better than the baselines (Tahmasebi, Borin, and Jatowt 2021).

2.2.1 Short-term Change

As mentioned in Section 1, some studies focus on investigating short-term meaning change, mostly employing textual data from social networks. Stewart et al. (2017) present a study on short-term change during the Russia-Ukraine crisis of 2014-2015, through data from VKontakte, a popular social media in the area. The aim of the research is to visualize and predict change in a word's semantics over the weeks, leveraging distributional representations. First, the tf-idf score for each word is extracted and concatenated in a time series that represents a concept drift, a measure which the authors define as a combination of a word's change in meaning and frequency. Then, temporal word embeddings are learned with the gensim implementation of word2vec: the vectors are initialized with the vocabulary of all the words in the data above a fixed frequency threshold, and then trained with tokenized posts for each weekly timestamp to generate a time series. Different word vectors are compared using cosine similarity and by looking at their neighbors. According to the authors, this study provides a generalizable proof of concept for future studies on short-term shift in social media.

Del Tredici, Fernández, and Boleda (2019) present an exploration of meaning shift within a period of 8 years with data from online community of speakers (sports subreddits), which allows better observation of short-term meaning shift. Previous research by Del Tredici and Fernández (2018) showed that this and similar communities have features that favor linguistic innovation. The behavior of a standard distributional model is tested when applied to short-term shift, showing that the model is confused by contextual changes due to particular references to people and event. A large sample of community-independent language is used to initialize the word vectors; then, these representations are updated for a certain point in time with the subreddit data.

Gonen et al. (2020) propose an alternative method to detect usage change. Specifically, they propose to work in the shared vocabulary space with the underlying intuition that words whose usage has changed are likely to be interchangeable with different sets

of words. Thus, these words will have different neighbors in the embeddings spaces of the two time periods. Their algorithm first represents each word in a corpus as the set of its top k nearest neighbors; then, it computes the score for word usage change across corpora by considering the size of the intersection of the two sets of neighbors. This method will be further discussed in Section 3.3., as it is the method that we selected for our experiment.

Guo, Xypolopoulos, and Vazirgiannis (2022) apply the method proposed by Hamilton, Leskovec, and Jurafsky (2016) to a corpus of tweets posted between April and June 2020. They compute *word2vec* word embeddings for each month, using the pre-trained twitter-200 gensim[12] model as reference. Then, they align the three obtained vector spaces to track usage change, and they present four case studies: *racism*, *hero*, *quarantine*, *ai*. Even in this small sample, the shift towards words related with COVID and healthcare is tangible: *racism* shifts away from *sexism* and *homophobia* towards *asians* and *sinophobia*. *hero* moves from *veteran* and *superman* towards *frontliner* and *covidwarrior*. *quarantine* goes from *swineflu* and *flu* to *coranatine* and *corona*. *ai* moves away from *math* and *data*, towards *ehealth* and *bloodtesting*. The authors also computed the stability distribution of words between the pre-COVID-19 reference and each of the three COVID-19 models, taking the average value as its final stability measure. They conclude that the meaning change across corpora is more significant than that over monthly time periods. To the best of our knowledge, this is the only study on English that has similar objectives to ours. However, there are some methodological differences: i. they follow the alignment approach of Hamilton, Leskovec, and Jurafsky (2016) to tracking usage change; ii. they focus on an arbitrarily selected group of key-words. For this reason the results are not fully comparable to ours.

2.2.2 Works on Italian

While the majority of the experiments on meaning change detection focuses on English, there are also studies for other languages. For example, SemEval 2020 Task 1 (Schlechtweg et al. 2020) addresses the unsupervised detection of meaning change in text corpora of German, English, Latin and Swedish.

As for Italian, some research has been conducted and presented at EVALITA 2020, under the DIACR-Ita: Diachronic Lexical Semantics task (Basile et al. 2020b). The corpus from Basile et al. (2020a), divided into two sub-corpora for the years 1945-1970 and 1990-2014, was used for the DIACR-Ita task. Several methods were submitted: Post-alignment, Joint Alignment, Contextual Embeddings, Graph-based and PoS tag features. Post-alignment systems first train static embeddings and then align them, while Joint alignment does these two processes at the same time. Contextual embeddings systems are based on contextualized embeddings, such as BERT (Devlin et al. 2019). Graph-based systems rely on graph algorithms, while PoS tag features systems use the distribution of targets PoS tags across the time slices. The majority of these systems use cosine distance as a measure of meaning change, except for Contextual embedding representations and Graph-based methods (Basile et al. 2020b). The best methods (Prazák, Pribán, and Taylor 2020; Kaiser, Schlechtweg, and im Walde 2020) use Skip-Gram with Negative Sampling (SGNS) to compute word embeddings, which are then aligned. Cosine similarity and a threshold are used to detect changed words.

Basile et al. (2016) employ Temporal Random Indexing, an embedding method first used in Basile, Caputo, and Semeraro (2014). The dataset is the Italian portion of

12 https://radimrehurek.com/gensim/

the Google Books Ngrams corpus, split into 10-year period sub-corpora for the time between 1850 and 2012. The vocabulary of each split vocabulary is modeled as the sum of its random vectors and then normalized to give less weight to the most frequent words. To detect shifts, the method by Kulkarni et al. (2014) is used. The study also employs temporal indexing to detect the average span of change in years (Tahmasebi, Borin, and Jatowt 2021).

Cafagna, De Mattei, and Nissim (2020) study how words are used differently in two Italian newspapers with diverging political opinions, *La Repubblica* (left-leaning) and *Il Giornale* (right-leaning). They focus on synchronic change, but the methodology is still relevant to the study of short-term usage change. The embeddings are first trained on *La Repubblica* texts and then updated with those from *Il Giornale*. The measure of the shift that the same word has undergone is then computed. A value for the frequency and a combination of both frequency and shift measure is also calculated. Starting from a shared vocabulary, the study features a top-down analysis, concerned with the change affecting the most frequent words in both newspapers; and a bottom-up analysis, that observes how a single word's usage varies across the two spaces looking both at its embeddings and frequency. It is proposed that the most interesting cases are those whose relative frequency does not change much in the two datasets, but still exhibit a high degree of change.

3 Methodology

This Section illustrates the methodology of our study: the creation of the corpus (§3.1) and its preprocessing (§3.2), as well as the details of the usage change algorithm we employed (§3.3).

3.1 Corpus

The dataset for this study is a newly created corpus of texts taken from Reddit.[13], a large on-line community made by more than 2.5 million user-created sub-communities called subreddits or subs.[14] As of December 2020, Reddit was the 18th-most visited website in the world, but it is still a mainly American phenomenon, with 41% of its traffic coming from the US,[15] where it is the 5th-most visited site.[16]. However, an active Italian community is present and is aggregated in a subreddit called *r/italy* from which we downloaded the texts composing our dataset.[17] Reddit gives free and easy access to historical data thus we were able to download posts (also known as submissions) and comments in the same specific time frame for the years 2019 and 2020, that is between January 30 and November 30. January 30 was chosen as the starting date of our period of interest because in 2020 it was the day when the first cases of COVID-19 were recorded in Italy. On the basis of these two time frames, the corpus is divided in two sub-corpora, one for each year (2019 and 2020).

We automatically built the corpus using a new Python 3 scraper script that allows accessing subreddit data through the Reddit API (Application Programming Interface). The Python implementation used in the scraper script is called PRAW (Python Reddit

13 https://www.reddit.com/
14 https://frontpagemetrics.com/history (as of December 2020).
15 https://www.alexa.com/siteinfo/reddit.com
16 https://www.redditinc.com/press (Retrieved December 30, 2020)
17 As of April 2021, *r/italy* had 300,000 subscribers.

API Wrapper).[18] However, using PRAW it is not possible to download posts or comments older than the last 1000 due to limitations in the Reddit API. To overcome this limitation, another API wrapper, called PSAW (Python Pushshift.io API wrapper), was used on top of the standard one.[19] This API leverages the pushshift.io[20] database for comment and submission search. Pushshift.io is a big-data storage and analytics project which copies data and metadata when they are posted on Reddit. The project also hosts monthly dumps of comments and submissions. These features make this project very useful for analyzing large quantities of Reddit data and, crucially, allows for the retrieval of data for a specific time range.

Our script was run two times, one for each time span. The script proceeds in the following manner: it first retrieves from Pushshift.io the IDs of all submissions in *r/italy* from the newest to the oldest; it then uses PRAW to collect the title of the submission, its text, and comments. A typical submission includes a title and a more articulated text; the discussion in the comment section is nested, as every user can answer to each comment. During the scraping, the raw text is iteratively saved in a text file for each day of the time frame; the texts are organized in two symmetrical folders. To ensure anonymity, no metadata regarding the author of the submission or comment is requested or saved in any way.

Despite being a very useful resource, that is, the basis for one of the few studies on Italian and the pandemic, and the first focusing on short-term usage change, the corpus we created has some limitations:

- Multiple languages: the majority of the downloaded texts are written in Italian, however there are some posts and comments in English. These tend to occur in the same context and submissions: most of them are posts from non-Italian users which are answered and discussed in English.
- Representativeness: as a 2016 American study showed, it should be noted that, as a whole, the userbase of Reddit is not representative of the overall population. Users of Reddit were once described as "offbeat, quirky, and anti-establishment".[21] This skewed demographic characterises also the number of users of *r/italy*: the userbase of the subreddit at the moment of the creation of the corpus was of 267,306 users, which is the 0.45% of the Italian population.[22]
- Accuracy of the texts: the Pushshift.io project API copies the submission at the moment of its creation on Reddit and does not update it. However, users often modify their posts and comments. These so called "EDITs" can be quite long and elaborate at times, and thus their absence may mean some loss of useful data. Moreover, some duplicate texts are present even if some specific restrictions were included based on the structure of Reddit

18 https://github.com/praw-dev/praw
19 https://github.com/dmarx/psaw
20 https://pushshift.io/
21 https://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/;
https://www.nbcnews.com/tech/tech-news/hipster-internet-favorite-reddit-may-have-lose-its-edge-go-n824866
22 The total population of Italy as of 1 January 2019 was 59,641,488 inhabitants (http://dati.istat.it/Index.aspx?QueryId=18460).

discussions. For instance, text of stickied posts[23] and comments are copied only once.

Despite these limitations, the resulting corpus has proved very useful to our purpose of detecting word usage change. Table 1 gives the size of the corpus and its subcorpora, both in terms of number of raw tokens and unique lemmas.

Table 1
Statistics about the corpus. The first column reports the number of white space-separated entities in the raw text files, the second column the number of unique tokens in the lemmatized text.

Year	Days	Raw Tokens	N. of Lemmas	Size
2019	305	24,141,080	283,570	151MB
2020	306	39,728,203	380,146	250MB

3.2 Pre-processing

We pre-processed the corpus following a two-step procedure: first we cleaned the texts and then we performed tokenization and lemmatization.

More specifically, the first step consisted in lowercasing all texts and removing URLs, special characters and stopwords.[24] Some special characters, however, were not removed in order to preserve specific features of Reddit, such as the use of / in the names of users and subreddit names, or the sarcasm tag /s. Words longer than 24 characters were substituted with the label "LONG" and a double paragraph break was added every 3,000 words to ease computation.

The second pre-processing step involved tokenization and lemmatization using Stanza (Qi et al. 2020). We chose Stanza because of its good performances on Italian texts: in particular, UD_Italian-ISDT is the model for which the highest accuracy is reported (97.79% for tokenization and 98.01% for lemmatization) compared to the other available models for Italian.[25] We performed lemmatization because it allows to focus on lexical meaning, removing morphological variations. The task also provided some interesting insights on problems that arise when lemmatization is applied to morphologically fusional languages, such as Italian. Indeed, a manual inspection of the processed data revealed that the lemmatizer, while performing well in the majority of the instances, had some problems with relatively uncommon words, borrowings, verbal forms, and named entities (e.g. names of states and nationalities or proper nouns and surnames). Other errors are due to non-standard spelling and form, or are the result of imprecise tokenization. Moreover, having employed an Italian lemmatization model, English words are not properly managed, for example they are often lemmatized by using Italian forms (e.g. vaccine, lemmatized as **vaccina*).

To get a rough estimation of the lemmatization quality, we compared a subsample of the lemmatized text against a list of valid Italian word-lemma pairs. After scoring 10 subsamples, we observed an average accuracy of 86%, well below the reported

23 That is, post and comments that are fixed in place by the subreddit moderators at the top of the page.
24 We used the Italian stop words of NLTK (https://www.nltk.org/).
25 https://stanfordnlp.github.io/stanza/performance.html

performances. We also repeated the experiment using SpaCy[26] achieving the same accuracy. We tried to improve the quality of lemmatization in several ways, for example by adding more stopwords, removing rare words (i.e. with less than 5 occurrences), or by normalizing the spelling of words to their most frequent form. However, these attempts did not result in a significant improvement of the final results. These additional experiments confirmed that lemmatization is a complex task when applied to morphologically complex languages. The problems are even more evident when dealing with noisy non-standard texts, such as spontaneous social network conversations, for which even state-of-the-art models, which in our case are trained mostly on news corpora, struggle to cope with.

3.3 Usage Change Detection

We adopted the method from Gonen et al. (2020), introduced in §2.2.1, to detect usage change. This method is perfectly in line with the aim of our study, that is to analyze differences between corpora by detecting words that are used differently across them. The task is defined by the authors as follows: given two corpora with substantial overlapping vocabularies, identify candidate words whose predominant use is different in the two corpora. The expected result is a ranked list of words, from the one that is most likely to have changed, to the least likely.

In other words, Gonen et al. (2020) propose to work in the shared vocabulary space with the underlying intuition that words whose usage changed are likely to be interchangeable with different sets of words, and so to have different neighbors in the two embedding spaces. Their algorithm represents each word in a corpus as the set of its top k nearest neighbors. Then, it computes the score for word usage change across corpora by considering the size of the intersection of the two sets.

Words with a smaller intersection are ranked higher as candidates for usage change. It is important to note that this method only considers the words in the intersection of both vocabularies, as words that are rare in one of the corpora are easily spotted by using their frequency in the two spaces, and do not fit the definition of usage change according to the authors. This method does not require extensive filtering of words; they instead filter words based on frequency, using a large value of $k = 1000$[4], because large neighbor sets are more stable.

The advantages of this method are plenty: (i) simplicity, since there is no need for space alignment, hyperparameter tuning and vocabulary filtering; (ii) interpretability, provided by the intuitive ranking system used for providing the results; (iii) locality, with the score for each word determined only by its own neighbors (whereas in the projection methods the similarity depends on the projection itself, which implicitly takes into account all the other words and their relations); (iv) stability, because the method produces similar results across different embeddings trained on the same corpora (this is not the case for alignment-based approaches). As the authors note, however, their method still has some limitations: it assumes high quality embeddings, and so, a large corpus. This is somewhat mitigated by the fact that the minimal input required is raw text without the need of annotation. In fact these are just minimal requirements; as already mentioned in §3.2, we used lemmatized text, where each token was substituted for its lemma. Another limitation is the fact that like previous approaches, this method does not guarantee that the detected words have indeed undergone usage change but

26 https://spacy.io/

it at least aims to highlight candidates for later human verification and interpretation (Gonen et al. 2020).

To adapt this method to our Italian corpus we firstly collated all the text of the two sub-corpora in two different text files, one for each year. The algorithm was then applied to these files, both in unlemmatized and lemmatized form, without altering any of its original parameters. The algorithm then computed the *word2vec* embeddings for both input files and returned the list of top-100 words which most likely have undergone usage change. Despite the imperfect results of lemmatization, we decided to focus our analysis on lemmatized text in order to reduce the problems connected to data sparsity and the morphological complexity of the Italian language. To visualize the output, we used t-SNE (t-distributed stochastic neighbor embeddings) as implemented in the method provided by Gonen et al. (2020) for the top-10 candidates, but we scaled down the number of represented neighbors of each word to enhance readability (see figure 1 as an example of the visualization).

4 Results and Discussion

In this Section we present and analyses the results obtained by the application of the usage change detection algorithm to the two sub-corpora.

4.1 Top-10 detected words

Table 2 lists all the top-10 neighbors in the 2019 and 2020 vector spaces for the top-10 candidate words detected by applying the Gonen et al. (2020) algorithm to our data.

The top-10 candidates can be divided in three broadly defined classes according to their nearest neighbors, and to how they have changed between the two sub-corpora. The fist class, *narrowing*, denotes candidates which changed from a more general usage, to a more specific one, but which is already present in the language. This is the case with *positivo*, *intensivo*, *guarire*, *gene*. The second class, *shift*, refers to those candidates which usage switched between two different semantic fields. Candidates such as *virus*, *testare*, *influenza* fall under this category. The last class, *not informative*, comprises those candidates which neighbors in both corpora, either due to their low frequency or noise, do not allow for a clear indication of usage. *bla*, *eco*, and *leve* are examples of not informative candidates proposed by the algorithm.

The word *positivo* (“positive”) is the first on the list. In the 2019 sub-corpus this adjective occurs mainly with terms pertaining to subjective evaluation (e.g. *recensione* “review” or *gradevole* “pleasant”) and emotional states (e.g. *ottimista* “optimist” and *attitudine* “aptitude”). The neighbors point to a meaning of *positivo* described in dictionaries[27] as usually employed in everyday language: "in an optimistic manner, with confidence, affirming the value of something or someone, good and favorable".

In the 2020 dataset, *positivo* has indeed narrowed its use to the medical semantic field: 8 out of its top-10 nearest neighbors are clearly connected with medicine and the pandemic. *tampone* (“swab”), *positività* (“positivity”), *40ena* (an abbreviation of *quarantena*, “quarantine”), *contagiare* (“to infect”), *sintomatico* (“symptomatic”), *asintomatico* (“asymptomatic”), [test] *sierologico* (“antibodies test”) and *infetto* (“infected”) all indicate a meaning of *positivo* as pertaining to medicine: a diagnostic response that confirms

27 The definitions of word senses in this section are taken from the online dictionary Treccani, https://www.treccani.it/.

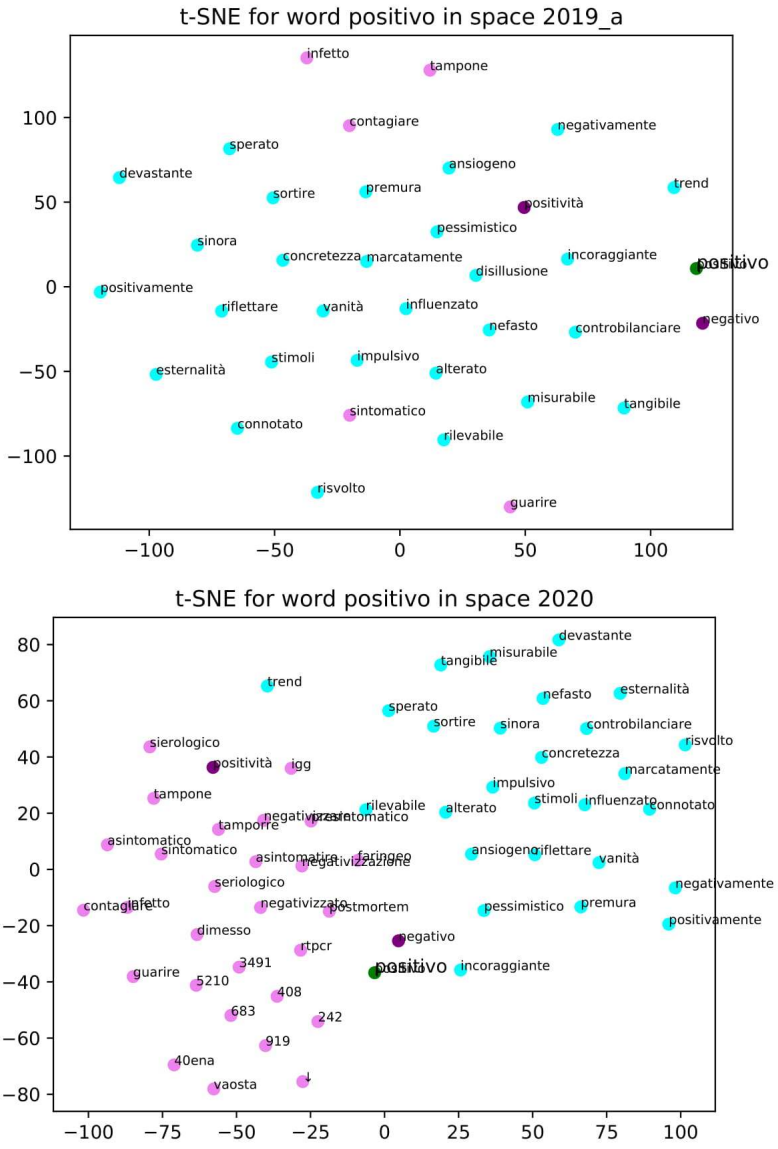

Figure 1
Visualization of *positivo* in the two sub-corpora

the formulated hypothesis, unfavorable to the tested subject, who, by extension, is also called *positivo*.

The other two nearest neighbors (NN) of *positivo*, *vaosta* and ↓, are less interpretable, however, can still be connected to the pandemic: the former is a shortening of Valle d'Aosta ("Aosta Valley"), the smallest Italian region, bordering France. *vaosta* occurs with the names of other regions in the daily tables listing COVID-19 cases and deaths. As to why only *vaosta* figures as a neighbor of *positivo* there is no evident clue. The symbol ↓ is present for the same reason: it occurs frequently in the periodic pandemic

reports. The connection with the tallies of the pandemic is confirmed by the presence of numbers in the neighborhood of *positivo*.

Figure 1 gives the visual representation of the top-30 nearest neighbors of *positivo* in the two spaces as an example of the plots created by the detection algorithm. Cyan is used for the 2019 neighbors, while pink is used for the 2020 ones. Shared neighbors are marked in purple. As can be seen for the visual representation, the only neighbors in common between the two sub-corpora are *positività* and *negativo*, marked in purple in the plot. The collocations of these two words in the 2019 sub-corpus reveal that they are almost never used in a medical sense.

The usage of *intensivo* is also characterized by a narrowing. In the first dataset, *intensivo* occurs in expressions like *corso intentensivo* ("crash course") or *allevamento intensivo* ("intensive animal farming"). Its neighbors, however, are diverse: *vegetale* ("plant, plant-related") is used both in talking about agriculture and food. *integratore* ("nutritional supplement") and *proteina* ("protein") are found in sentences about *allenamento intensivo* ("intensive training"). *cbd* (cannabidiol), *thc* (tetrahydrocannabinol),[28] *molecola* ("molecule") and *nicotina* ("nicotine") are related to drugs. These could well be connected to *combustione* ("combustion"). The connection with *intestino* ("intestine") and *hiv* (human immunodeficiency viruses) are less clear. However, after looking at the 2020 sub-corpus, it is clear that the usage of *intensivo* in the 2019 dataset was more general.

Indeed, in the 2020 sub-corpus, the word *intensivo* is used more frequently, and all its neighbors belong to the healthcare vocabulary. *icu* (intensive care unit) and its Italian counterpart *rianimazione* ("reanimation", used interchangeably with *terapia intensiva*) are the firsts on the list. Further down, there are *ricovero/ospedalizzazione* ("hospitalization"), *ricoverare/ospedalizzare* ("to hospitalize"), *ricoverato/ospedalizzato* ("hospitalized"), and *ospedale* ("hospital"). *intubare* refers to the operation performed by doctors to install a breathing tube into the throat of a patient. Here the connection to the pandemic is pervasive, with some neighbors hinting at the strenuous conditions of hospitals during the pandemic (*saturato* "saturated"), and other names of Italian regions (*vaosta*,[29] *friulivg* (Friuli-Venezia Giulia)).

The usage of *guarire* ("to heal") also underwent narrowing, from a broader use in the medical semantic field, to a more restricted use regarding intensive care. In the 2019 space, the neighbors of *guarire* feature words such as *psicoterapeuta* ("psychotherapist"), *deprimere*, ("to depress") and *psicologo* ("psychologist") which pertain to mental health and therapy. *astinenza* ("abstinence", or "withdrawal") and *malessere* ("discomfort") can also be connected to this scope. *chirurgo* ("surgeon"), *dottoressa* ("female doctor") and *prescrivere* ("to prescribe") are more general. *intestino* and *involontariamente* ("unintentionally") seem incidental and are not particularly informative.

28 CBD and THC are two of the cannabinoids found in cannabis.
29 A wrongly lemmatized **vaosto* is present

Table 2
Neighbors of top-10 words. The upper line lists top-10 neighbors in the 2019 sub-corpus, while the lower line lists neighbors in the 2020 sub-corpus. Lemmatization errors, which are almost always easily understandable by a native speaker, are marked with a star (*) symbol.

	Neighbors
positivo	desiderato, decisivo, ottimista, lato, recensione, gradevole, normalità, attitudine, scaturire, deleterio
	tampone, positività, 40ena, contagiare, sintomatico, vaosta, asintomatico, ↓, sierologico, infetto
virus	vulnerabilità, bios, resettare, bug, chiavetta, terminale, criptare, diabete, cancro, scansione
	sarscov2, coronavirus, covid19, covid, ebola, contagio, patogeno, infettare, sars, aerosol
intensivo	vegetale, integratore, combustione, cbd, thc, molecola, nicotina, proteina, intestino, hiv
	icu, rianimazione, ospedalizzato, ricovero, ricoverare, ricoverato, ospedalizzare, intubare, ospedale, ospedalizzazione
testare	arduino, centralina, plugin, usato, apportare, alterare, simulare, falla, fungere, lsd
	tampone, sintomatico, testato, asintomatico, ospedalizzare, ct, diagnostico, tamponare, screening, infetto
guarire	psicoterapeuta, astinenza, prescrivere, chirurgo, dottoressa, intestino, deprimere, malessere, psicologo, involontariamente
	guarito, ospedalizzare, asintomatico, guarigione, clinicamente, ricoverato, contagiare, decesso, decedere, ospedalizzato
bla	professorone, la', sminuire, inerzia, passere, *ripartare, umiltà, contrattuale, pigrizia, *dirtelare
	superfluo, trito, figliare, *diciamocelare, egoistico, famigliare, etc, *smettilare, moralista, stufo
eco	intitolare, convegno, studentesco, seguace, sostenitore, invocare, *buongiorne, vocabolario, *altaforte, rivista
	asimov, pascolo, collana, lovecraft, mattone, divulgativo, *murakamo, microscopico, orwell, philip
leve	*capacitare, sara, immortale, magnetico, aggrappare, rivoluzionario, cosmico, *vadere, lentezza, nano
	inculare, perverso, rum, ftw, dinamico, trucchetto, proletariato, arrampicato, toppa, nervo
influenza	interferire, migratore, venezuelano, decisivo, presidenziale, caratterizzare, coinvolgimento, connotazione, competitività, oppressione
	influenzale, *polmonito, stagionale, sintomatologia, *polmonita, complicanza, ebola, mers, morbillo, contagiosità
gene	amministrare, neurone, tribù, bravura, portatore, risaputo, composto, azionista, squilibrio, prole
	dna, nomea, ariano, innato, land, sarscov2, mutazione, ceppo, cromosoma, mediocrità

In the 2020 sub-corpus, the neighborhood of *guarire* is focused on hospital and intensive care: *ospedalizzare*, *ospedalizzato*, *ricoverato*, *contagiare*, and *asintomatico* return as neighbors. These words are common also in the surroundings of other pandemic-related words. Other terms are both positive, like *guarito* ("healed") and *guarigione* ("healing"), and negative, like *decesso* ("death") and *decedere* ("to die"). These last two are commonly used in a more formal setting, like news, so are likely connected to *guarire* because of the reports on cases, deaths, and recoveries from COVID-19. This is confirmed by the presence of numbers in the larger neighborhood. *clinicamente* ("clinically") is more neutral, but still connected to the field of medicine. The only common neighbor between the two spaces is *guarigione*.

gene is the last detected word in the top-10 candidates which could be ascribed to the *narrowing* class. The low number of occurrences does not allow for a coherent embedding representation in both spaces, even if the 2020 one seems slightly better than the one from the previous year's dataset. While the neighborhood in the 2019 space is quite diverse, in the 2020 one can identify some connections: *dna*, *ariano* ("arian"), *mutazione* ("mutation"), *ceppo* ("strain", meaning a variant of e.g. a virus), *cromosoma* ("chromosome") relate to the genetic sense of *gene*. *nomea* ("reputation"), *innato* ("innate"), and by contrast *mediocrità* ("mediocrity"), loosely hint at sense of "genius". However, the influence of the pandemic may have prompted more discussion that involved the biological gene: *sarscov2* is listed as a neighbor in the 2020 space. Interestingly, there are no shared words in the neighborhoods for the two spaces.

The usage of *virus* shifted from informatics to the pandemic. Eight out of ten neighbors in 2019 point to the computer version of a virus: *vulnerabilità* ("vulnerability" of a system), *bios*,[30] *resettare* ("to reset"), *bug*,[31] *chiavetta* ("USB pen-drive"), *terminale* ("terminal"), *criptare* ("to encrypt"), and *scansione* ("scan"). On the contrary, *diabete* ("diabetes") and *cancro* ("cancer") are medical terms.

In the 2020 sub-corpus, the connection with the pandemic is explicit in the first four neighbors, which are all variations of COVID-19: *sarscov2*, *coronavirus*, *covid19* and *covid*. The other neighbors are still correlated with specific diseases, like *ebola* and *sars*, and their spread, *contagio* ("contagion"), *patogeno* ("pathogen"), *infettare*. *aerosol* refers to the fact that COVID-19 is spread by droplets of saliva in the air. Even looking at a larger list of neighbors, there are no words related to the computer sense of *virus* during 2020. Common neighbors between the two corpora are *infettare*, *contagioso*, *infezione*, *hiv*.

Also the usage of *testare* has shifted from practical engineering to medical testing. Some neighbors of *testare* ("to test") in the 2019 sub-corpus hint at electronic and digital devices: *arduino*,[32] *centralina* ("control unit"), *plugin*, all belong to this semantic field. Other neighbors include the verbs *alterare* ("to alter"), *simulare* ("to simulate") and *fungere* ("to function"), which are used in diverse situations. *usato* ("used"), *falla* ("fault") and *lsd* (lysergic acid diethylamide) range from generic to very specific. As it was the case with *intensivo*, again the sparse use of this word in the 2019 dataset renders its representation imprecise, overly influenced by usage in only certain limited discussions.

30 Acronym for Basic Input/Output System, a firmware (a special kind of software that provides low-level control for a specific hardware) used to perform hardware initialization during the power-on startup process, known as booting.

31 A software bug is an error, flaw, or fault in a computer program that causes an incorrect or unexpected result or behavior. This is indeed the sense of bug here, as it is common to use English loanwords or calques in Italian for the digital semantic field in general.

32 Arduino is an open-source hardware and software company project and user community that designs and manufactures single-board microcontrollers and microcontroller kits for building digital devices.

In the 2020 sub-corpus, *testare* is well represented and connected to the pandemic: *tamponare* arises as a specialized version of *testare*, with the specific meaning of "to test with a swab (*tampone*)". This is connected to *screening*, a loanword that refers to medical testing. The acronym *ct* is used interchangeably for COVID-19 test and *commissario tecnico* ("sports coach"). *diagnostico* ("diagnostic") seems to be used in connection with *tampone* and other testing methods and equipment, confirming the connection to the pandemic. *sintomatico* ("symptomatic"), *asintomatico* ("asymptomatic"), *infetto*, and *ospedalizzare* are clearly connected to the virus.

tamponare is present in the list also as a neighbor of *isolare*, which is discussed in Section 4.2. This is an interesting case: this meaning of *tamponare* as "to perform a swab" is listed as a 2020 neologism derived from *tampone* ("swab") on the on-line version of the Treccani dictionary. This is supported by the data in our corpus: its nearest neighbors in the 2020 space are words such as *testare*, *sintomatico*, *malauguratamente* ("unfortunately"), *tampone*, *ospedalizzare*, and *profilassi* ("prophylaxis"). Thus, this lemma is an homonym of *tamponare* in the sense of "to hit, with the anterior part of a vehicle, the back of another vehicle in the same lane". The neighbors of *tamponare* in the 2019, point only at this sense: *cid*, abbreviation for "Convenzione d'Indennizzo Diretto, or Constatazione Amichevole d'Incidente Stradale" (lit. "Direct Compensation Convention" or "Friendly Verification of Car Accident"),[33] clearly pertains to the car accident situation. Other neighbors in 2019 are *bagagliaio* ("trunk"), *sopraggiungere* ("to arrive, usually suddenly and unexpectedly"), *retromarcia* ("reverse [gear]"), *frenata* ("hard braking"), *frenare* ("to brake"), *conducente* ("driver"), *semaforo* ("traffic light").

The word *influenza* has shifted its use, too. Its neighbors in the 2019 sub-corpus point at the sense of "action done by one thing or person on another one": *interferire* ("to interfere"), *coinvolgimento* ("involvement"), and *caratterizzare* ("to characterize") are quite indicative in this sense. Some other hint at a more geo-political use of the same sense of *influenza*: *presidenziale* ("presidential"), *venezuelano* ("Venezuelan"), *oppressione* ("oppression"), and **migratore* (maybe *migratoria* "migratory", as in *flussi migratori*). Collocations shows that *presidenziale* is used referring to American politics, while *venezuelano* refers to the Venezuelan crisis in the beginning of 2019. *connotazione* ("connotation"), *competitività* ("competitivity") and *decisivo* ("decisive") are the remaining neighbors.

In the 2020 sub-corpus, *influenza* shifts completely to a medical usage: some of its neighbors refer to diseases, such as *ebola*, *mers*, *morbillo*, **polmonito* and **polmonita* (correct form: *polmonite*, "pneumonia"). *stagionale* is coming from *influenza stagionale* ("common flu"), while *influenzale* ("flu-related"), *sintomatologia* ("symptomatology"), *complicanza* ("complication"), and *contagiosità* ("the ability or state to be infective") relate to the effect of *influenza*. Even in the larger neighborhood there is no trace of the usages attested in 2019. Moreover, there are no neighbors in common between the two datasets.

bla is the first of the three candidates which neighbors are not informative with respect to usage change. *bla*, usually repeated two or three times (*bla bla*), is a common onomatopoeia indicating useless conversations or futile chatter. The frequency of this word grew only slightly, in line with the overall increment in the size of the data. In both spaces the neighbors of *bla* are not informative. The only ones that can be somewhat connected with the common use of *bla* are found in the 2020 space: *trito* ("crushed"), can be used in the idiomatic expression *trito e ritrito* ("grounded and grounded again") meaning something that is used or said too much, commonly known, prosaic and trivial.

33 This is referring both to a procedure and its related form that allows for more smooth insurance compensation of the damage.

Table 3
Absolute and relative frequencies of top-10 words. Relative frequencies are calculated as the number of occurrences of a word divided by the total number of tokens in the lemmatized corpus for a specific year.

	Absolute Frequency		**Relative Frequency (%)**		**Increase (%)**
	2019	**2020**	**2019**	**2020**	
positivo	2969	13688	0.52169	1.78806	1.26637
virus	270	16632	0.04744	2.17263	2.12519
intensivo	203	3716	0.03567	0.48542	0.44975
testare	439	3347	0.07714	0.43722	0.36008
guarire	238	2454	0.04182	0.32057	0.27875
bla	302	489	0.05307	0.06388	0.01081
eco	256	436	0.04498	0.05695	0.01197
leve	269	369	0.04727	0.04820	0.00093
influenza	996	4226	0.17501	0.55204	0.37703
gene	354	524	0.06220	0.06845	0.00625

Others are *superfluo* (“excessive”) and *etc* (abbreviation of “etcetera”). The neighbors show some lemmatization issues: **ripartare* instead of *ripartire* (“to start again”), **dirtelare* most certainly derived from *dirtelo* (“to say to you”, *-lo* is an enclitic second person pronoun), **diciamocelare* from *diciamocelo* (“to say to ourselves”, often said with the sense of “let’s be clear to/real with ourseves”) and **smettilare* from *smettila* (“stop it!”).

Also not very informative are the neighbors of *eco*, which can indeed be the common noun for “echo”, a reflection of sound; or, if one looks at its nearest neighbors, the famous writer Umberto Eco, at least in the 2020 sub-corpus. However, in the 2019 dataset the neighborhood is less clear even if they are clearly related to literature: for example, *pascolo* referring to poet Giovanni Pascoli and *murakamo* referring to Japanese writer Murakami Ryū. Other words connected with the literary world are *collana*, a series of books, *mattone*, a long and tedious book (lit. “a brick”), and *divulgativo*, usually a science or otherwise academic book intended for the general audience. *philiph* is referring to sci-fi author Philip K. Dick, as suggested by the presence of other writers of the same genre like Isaac Asimov (*asimov*) and George Orwell (*orwell*).

leve is maybe the least informative entry in this list: it is a lemmatization error, since the lemmatizer did not use the citation form, the singular *leva* (“lever”). In addition, in some cases *leve* can derive from the name of Holocaust survivor and writer Primo Levi. The neighborhood of *leve* in both corpora emerges from limited interactions in peculiar discussions: just to cite one, the first neighbor of the 2020 list, a profanity that literally means “to sodomize”, is due to an exchange between two users on day 261 of 2020, where the word *leve* was used about 20 times.

The presence of these three cases, *eco*, *bla*, *leve*, can be attributed to their inaccurate embedding representations, which are in turn due to their scarce frequency. Their word embeddings are overly influenced by some peculiar context of use, which renders their neighborhoods less informative to define their usage. Frequency-wise all three share a pattern of just a slight increase, in line with the growth of the data for the second corpus. A preliminary analysis of the other results in the top-100 detected words shows that this can be the case for many other relatively low-frequency words.

Overall, in these top-10 candidates there are 6 informative results (*positivo*, *virus*, *intensivo*, *testare*, *guarire*, *influenza*, *gene*), and 3 less informative results (*bla*, *eco*, *leve*). These last three candidates have imprecise embedding representations, due to their low frequency of use. Also, wrong lemmatization may have had an impact. However, among all the 200 neighbors of the words listed in the top-10, just 11 are wrongly lemmatized, and even in these cases the errors are intelligible for a native speaker.

Sometimes, the embedding representations of the top-10 candidates show some less specific neighbors, at least in the 2019 space, where their frequency is lower. It is interesting to note, however, that all the informative candidates had a noticeable growth in relative frequency in the 2020 sub-corpus. As seen with informative outputs, if the changed word is well represented, it is also detected by the algorithm. Table 3 gives frequency data for the top-10 candidate words proposed by the algorithm.

Even if the top-10 results are not totally devoid of problems, the output for an unlemmatized corpus seems worse, with only five terms (*virus*, *bla*, *vaccino*, *positivo*, and *positivi*) having some significant increase in frequency. As seen in the previously discussed words, low frequency leads to imprecise representations built only on a handful of particular discussions. This naturally leads to radically different neighborhoods over the two corpora, tricking the algorithm into thinking that these cases are instances of usage change. It can be argued that these words have in fact undergone usage change, that is, they have changed contexts of use, but their neighbors often give no clue to their meaning, calling into question their validity. While far from being perfect, lemmatization seems to smooth out at least some of these cases. Pertaining to specific results in the unlemmatized top-10, *bla* presents the same problems as explained above; *positivo* and *positivi* are two inflected forms of the same lemma (thus of low informative value when searching for changes in language use), but overall correctly labeled as changed; *peste* and *vaccino* have a clear enough representation only in the second sub-corpus; *fico* has sparse usage in both datasets, but it is clear only in the first one. *capitano* is the only case with decreased frequency: in the 2019 space it is related almost exclusively with the discussion around the incident involving NGO ship captain Carola Rackete and her antagonist, then Interior Minister Matteo Salvini, sometimes nicknamed "Il Capitano". In the 2020 space many neighbors point to *càpitano* as a verb ("they happened", as opposed to *capitàno*, "captain").

As for the top-10s detected with the method based on the alignment of the vector spaces (AlignCos, Hamilton, Leskovec, and Jurafsky (2016)), in both cases they are much worse than those found by the Nearest Neighbors method (Gonen et al. 2020): in the lemmatized version, only *intensivo* is significative, all the others being cases of representations skewed by low frequency. In the unlemmatized version two candidates are somewhat valid: *fontana* changed its use from "fountain" in the 2019 space to referring to Attilio Fontana, the governor of Lombardy, the Italian region hit the worst by the pandemic. The other significant candidate is *vaccino*, which has a low frequency in 2019 and an obviously good representation in 2020.

Table 4
Neighbors of other relevant words in the top-100. The upper line lists top-10 neighbors in the 2019 sub-corpus, while the lower line lists neighbors in the 2020 sub-corpus. Lemmatization errors, which are almost always easily understandable by a native speaker, are marked with a star (*) symbol.

	Neighbors
vaccino	omeopatico, prescrivere, biologo, dermatologo, ricoverare, cancro, esente, omosessualità, *asile, prescrizione
	antinfluenzale, antivirale, pfizer, morbillo, vaccinato, oxford, anticorpo, somministrare, cavia, gregge
riaprire	avviare, portone, archiviare, spostato, *accendare, riprovare, quirinale, rimbalzare, sbucare, avvio
	riapertura, richiudere, ripartenza, palestre, maggio, allentare, restrizione, *asile, allentamento, *contage
tappeto	muffa, vernice, soffitto, siringa, lavandino, cemento, ombrellone, rame, tavoletta, pallino
	testare, tampone, sierologico, screening, capillare, prericovero, isolare, quarantene, molecolare, test
normalità	maschilista, retrogrado, omofobia, immaturo, socialmente, bigotto, ansioso, omosessualità, geloso, deleterio
	riaprire, andata, allentamento, parvenza, gradualmente, autunno, riapertura, intimità, ricaduta, esodo
morto	persecuzione, attenuante, portatore, perseguitare, ignoto, terrorizzare, rapire, concentramento, stupratore, molestia
	ospedalizzato, decesso, contagiato, infettato, ricoverato, *contage, ospedalizzare, decedere, *muoiare, contagiare
curva	ultras, tifoso, tifoseria, vicolo, tir, marce, interista, lazio, hamilton, *filmetro
	esponenziale, appiattire, curvo, contage, accelerazione, pendenza, impennare, picco, r0, progressione
isolare	inadatto, interferire, sabotare, dovunque, volente, fomentare, deviare, bollare, nolente, quotidianità
	circoscrivere, *focolao, quarantenare, sintomatico, blindare, rintracciare, *diffondare, asintomatico, tamponare, contagio
ondata	antisemitismo, migratore, sovranismo, generazionale, retorica, buonismo, berlusconiano, reazionario, consumismo, apice
	epidemia, *focolao, impennata, lockdown, esodo, autunno, scongiurare, riapertura, pandemia, *tsuname
terapia	omeopatico, prevenzione, malessere, allergia, raffreddore, relazionale, stigma, erezione, molecola, ansioso
	rianimazione, ricovero, ospedalizzare, intubare, icu, ospedalizzato, farmacologico, ospedalizzazione, ricoverato, ormonale
rosso	guancia, muffa, romeo, cera, hamilton, mela, dannato, illuminare, adesivo, cappello
	zona, grigie, nembro, lodigiano, tonalità, *mantovo, tartufo, codogno, stemma, *cremono
scorta	saviano, divisa, domiciliare, proiettile, equipaggio, rinchiudere, digos, immunità, rimozione, vendicare
	rifornire, scarseggiare, *vivero, igienizzante, *amuchino, introvabile, assaltare, monouso, sottovuoto, ricambi
chiuso	finito, tappo, elefante, fogna, circolo, serratura, sbraitare, sfortunatamente, *tienire, cassetto
	blindare, palestre, richiudere, affollare, riapertura, blindato, battente, scappato, ammassare, confinare
fontana	basilica, sant', cimitero, sottosegretario, virginia, passeggiare, *lucco, riva, villa, portone
	*gallero, cirio, *zaio, gallera, *bonaccino, *formigone, assessore, *camico, *umilansifonsifere, lombardia
emergenza	abitativo, naufrago, irregolarità, *rimpatro, rifugiare, irregolare, incidente, interruzione, *lampeduso, generatore
	pandemia, emergenziale, pandemico, epidemia, fronteggiare, proroga, imprevista, commissariamento, ripartenza, *covere
paziente	lucidità, malessere, dolore, nutrizionista, ossessivo, allergia, seduta, ansioso, erezione, perizia
	rianimazione, ospedalizzare, intubare, 38enne, anestesista, complicanza, *polmonita, oncologico, ricoverato, diagnosticato
malato	frustrato, sindrome, isterico, risvegliare, disordine, omeopatico, ossessivo, immaturo, retrogrado, daenerys
	oncologico, ammalato, infetto, ospedalizzare, contagiato, diagnosticato, immunodepresso, *ammalere, ricoverato, contagiare

4.2 Other relevant words

Other words in the top-100 candidates for usage change are relevant. These and their neighbors are listed in Table 4 and are briefly discussed below.

Notable cases of usage narrowing include *vaccino*, *terapia* ("therapy"), *malato* ("ill"), and *paziente* ("patient"). The neighbors of these words change from generally mild connotation (e.g. *terapia* is used in the 2019 sub-corpus with *omeopatico* "homeopathy", *allergia* "allergy", and *raffreddore* "common cold"), to a more severe one, related to the pandemic (e.g. *terapia* is used in the 2020 sub-corpus with *rianimazione*, *intubare*, and *icu*).

Other instances of narrowing, or change to a specific usage, are those of *curva* ("curve"), *rosso* ("red"), and *isolare* ("to isolate"), which becomes to be specifically connected to the pandemic. Among the latter's neighbors *quarantenare* ("to quarantine") is found. The case of *quarantenare* is interesting, despite very low occurrences. This verb is found a dozen of times in the 2020 space, but only one in the 2019 one. It is also not present in the Treccani dictionary, not even as a neologism. *quarantenare* is used in the sense of "to quarantine". The 2019 instance refers to the process with which Reddit administrators hide and close a subreddit deemed to be harmful or not in line with the platform's rules. The sense in 2020 is similar, but the term is used always in connection with the pandemic: neighbors include *quarantena* (lemmatized as **quaranteno*), *fiduciario* ("fiduciary", in the more formal expression "quarantena fiduciaria"), *infettare* ("to infect"), *autoisolamento* ("self-isolation"), *isolamento* ("isolation"), *precauzionale* ("precautionary").

The neighbors of *normalità* ("normality"), *tappeto* ("carpet"), *morto* ("dead"), *ondata* ("wave"), *scorta* (both "security detail" and "stockpile"), *emergency* ("emergenza") point to shifts in usage. *normalità* shift its usage from gender issues to the pandemic; *tappeto* moves from homes to "carpet testing"; *morto*'s usage chages from crimes to pandemic deaths; *ondata* from a geopolitical usage to describing the successive waves of the disease; *scorta* switches from "security detail" to "stockpile"; and *emergenza* refocuses from the migrants crisis to the pandemic.

Table 5 gives frequency data on other relevant words in the top-100, The words are ordered according to their position in the list proposed by the algorithm (not always contiguous), as already seen for the top-10 candidates; words which have indeed experienced change have also a noticeable increase in relative frequency, albeit less than for the terms in the top-10.

5 Conclusions

This work started from the hypothesis that a global crisis like the COVID-19 pandemic could impact language use. This was verified with computational means, leveraging both theoretical linguistics and NLP techniques. A corpus was created by scraping online text from the Italian Reddit community. The data was collected for the days between January 30 and November 30 of both 2020 and 2019, creating two sub-corpora. The raw text was then cleaned and lemmatized to allow further analysis. This dataset alone, both raw and preprocessed, could be a useful resource for other applications and it is publicly available. Future work may focus on the extension of the dataset's timeframe.

This research follows previous work and methodology in the field of computational language change detection, focusing on short-term usage change. To the best of our knowledge, this is the first work of this kind done for Italian, both in the field of

Table 5
Absolute and relative frequencies of other relevant words in the top-100. Relative frequencies are calculated as the number of occurrences of a word divided by the total number of tokens in the lemmatized corpus for a specific year.

	Absolute Frequency		Relative Frequency (%)		Increase (%)
	2019	2020	2019	2020	
vaccino	710	5677	0.12476	0.74158	0.61683
riaprire	311	3775	0.05465	0.49313	0.43848
tappeto	226	787	0.03971	0.10281	0.06309
normalità	265	1185	0.04656	0.15480	0.10823
morto	2201	7426	0.38675	0.97006	0.58331
curva	692	1670	0.12159	0.21815	0.09656
isolare	240	994	0.04217	0.12985	0.08767
ondata	233	1750	0.04094	0.22860	0.18766
terapia	738	4857	0.12968	0.63447	0.50479
rosso	2791	6457	0.49042	0.84348	0.35306
scorta	441	1047	0.07750	0.13677	0.05928
chiuso	1856	7148	0.32613	0.93374	0.60762
fontana	253	1264	0.04446	0.16512	0.12066
emergenza	1131	5399	0.19873	0.70527	0.50654
paziente	873	4125	0.15340	0.53885	0.38545
malato	954	3289	0.16763	0.42964	0.26201

COVID-19-related linguistic research and short-term language change detection. The latter is carried out with the neighborhood-based method outlined in Gonen et al. (2020), previously untested for the Italian language. The choice to lemmatize the data allowed to evaluate the impact of this pre-processing step on the method.

The initial research questions were the following: has the pandemic impacted the usage of the Italian language? Can this impact be detected with computational means? In fact, the manual analysis of the results produced by the algorithm showed that, as expected, some degree of usage change has occurred. The computational method used to detect it has shown to be quite solid also for Italian. Our experiments have shown that lemmatization as a pre-processing phase is important for Italian, given that without this step the results were less informative, although it remains a challenging task for an inflectional language such as Italian. Future work may involve an improved lemmatization.

A similar work for English by Guo, Xypolopoulos, and Vazirgiannis (2022) adopts different methodological choices, such as the alignment approach of Hamilton, Leskovec, and Jurafsky (2016) to detect usage change, and the analysis of a selection of predefined keywords. Its results are therefore not comparable to ours. Nevertheless, it shows that, also for English, a shift in usage is detected towards COVID19 and healthcare related words.

It remains to be seen if the change in usage will translate in actual lasting mutations in the language. The rise of of a new word as in the case of *tamponare* "to perform a swab", may be more significant and enduring than the already existing, but domain-specific sense of *positivo* "a diagnostic response that confirms the formulated hypothesis, unfavorable to the tested subject", which became widespread due the pandemic. These

cases seem more typical of short-term usage change: more specific, or different senses of a word increase their use, and overtake the more established senses due to a plethora of factors, in this case the pandemic. The surge of these senses may well be temporary.

In conclusion, this work successfully contributed to: (i) the creation of a new dataset, focusing on short-term usage change for Italian; (ii) the cross-linguistic application of a relatively novel method of language change detection; (iii) a linguistic analysis of the impact of the pandemic on language use in a language other than English. All the data and part of the code created for this work are publicly available online.[34]

Acknowledgments

We would like to thank the reviewers for their valuable comments and suggestions.[35]

References

Bamler, Robert and Stephan Mandt. 2017. Dynamic word embeddings.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020a. A Diachronic Italian Corpus based on "L'Unità". In *Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it 2020)*, volume 2769, Online, March 1-3, 2021. Italian Association for Computational Linguistics.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020b. DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online, December.

Basile, Pierpaolo, Annalina Caputo, Roberta Luisi, and Giovanni Semeraro. 2016. Diachronic analysis of the italian language exploiting google ngram. In Pierpaolo Basile, Anna Corazza, Francesco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, volume 1749 of *CEUR Workshop Proceedings*, Napoli, Italy, December 5-7. CEUR-WS.org.

Basile, Pierpaolo, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Blank, Andreas. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. *Historical Semantics and Cognition*.

Cafagna, Michele, Lorenzo De Mattei, and Malvina Nissim. 2020. Embeddings-based detection of word use variation in italian newspapers. *Italian Journal of Computational Linguistics*, 6:9–22, 12.

Del Tredici, Marco and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Del Tredici, Marco, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota, June. Association for Computational Linguistics.

34 https://github.com/edoardosignoroni/usage_change_ITA

35 The paper is a re-elaboration of Edoardo Signoroni's Master Thesis defended at the University of Pavia. For the only purposes of the Italian Academia, Elisabetta Jezek is responsible for sections 1 and 2, and Edoardo Signoroni for sections 3, 4 and 5. Rachele Sprugnoli edited the paper before the submission.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Eger, Steffen and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 52–58, Berlin, Germany, August. Association for Computational Linguistics.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online, July. Association for Computational Linguistics.

Gulordava, Kristina and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July. Association for Computational Linguistics.

Guo, Yanzhu, Christos Xypolopoulos, and Michalis Vazirgiannis. 2022. How COVID-19 is Changing Our Language: Detecting Semantic Shift in Twitter Word Embeddings. In *Conférence Nationale en Intelligence Artificielle 2022 (CNIA 2022)*, Actes CNIA 2022, Saint-Etienne, France, June.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Harris, Zellig. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Ježek, Elisabetta. 2016. *The Lexicon, an Introduction*. Oxford Textbooks in Linguistics. Oxford University Press, Oxford.

Jurafsky, Dan and James H. Martin. 2021. *Speech and Natural Language Processing (3rd ed. draft)*. Pearson Prentice Hall. retrieved from https://web.stanford.edu/ jurafsky/slp3/.

Kahmann, Christian, Andreas Niekler, and Gerhard Heyer. 2017. Detecting and assessing contextual change in diachronic text documents using context volatility. In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 135–143, Funchal, Madeira, Portugal, November.

Kaiser, Jens, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online, December.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, June. Association for Computational Linguistics.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically significant detection of linguistic change.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–71.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Prazák, Ondrej, Pavel Pribán, and Stephen Taylor. 2020. UWB @ DIACR-Ita: Lexical Semantic Change Detection with CCA and Orthogonal Transformation. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online, December.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*,

Online, July.
Rodda, Martina A., Marco Senaldi, and Alessandro Lenci. 2017. Panta rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3:11–24, 06.
Rudolph, Maja and David Blei. 2018. Dynamic bernoulli embeddings for language evolution. In *Proceedings of The Web Conference 2018*, Lyon, France, April. ACM.
Sagi, Eyal, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece, March.
Sahlgren, Magnus. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–53.
Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online), December. International Committee for Computational Linguistics.
Stewart, Ian, Dustin Arendt, Eric Bell, and Svitlana Volkova. 2017. Measuring, predicting and visualizing short-term change in word representation and usage in vkontakte social network. In *Eleventh international AAAI conference on web and social media*, Montreal, Canada, March.
Tahmasebi, Nina, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, pages 1–91.
Tang, Xuri, Weiguang Qu, and Xiaohe Chen. 2013. Semantic change computation: A successive approach. In Longbing Cao, Hiroshi Motoda, Jaideep Srivastava, Ee-Peng Lim, Irwin King, Philip S. Yu, Wolfgang Nejdl, Guandong Xu, Gang Li, and Ya Zhang, editors, *Behavior and Social Computing*, pages 68–81, Cham. Springer International Publishing.
Tang, Xuri, Weiguang Qu, and Xiaohe Chen. 2016. Semantic change computation: A successive approach. *World Wide Web*, 19.
Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, Los Angeles, California, USA, February. ACM.