ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

The Economics of Obituaries

Raffaella Intinghero
Pietro Lazzaretto
Paolo Manasse

Quaderni - Working Paper DSE N°1209

The Economics of Obituaries*

Raffaella Intinghero†, Pietro Lazzaretto‡, Paolo Manasse§

August 18, 2025

Abstract

Obituaries are traditionally seen as expressions of grief and remembrance. We argue that they also have an underappreciated economic role: they are vehicles for strategic social and economic signaling. In this paper, we develop a simple theoretical framework in which paid obituaries serve as a form of self-promotion for the authors, especially when the deceased is a prominent public figure. We then test this hypothesis using data from Italy, exploiting the variation in mortality caused by the COVID-19 pandemic as a natural experiment. We show that higher mortality rates are associated with increases in per-capita obituaries, driven not by informational needs but by strategic advertising motives. Our results suggest that obituaries function as a marketplace for visibility and status, where social and economic incentives intersect.

Keywords: Economics of Obituaries, Status Signaling, Social Networks, Media and News, COVID-19, Text Analysis.

JEL Codes: A10, A13, D71, D85, D91.

*We sincerely thank Giulio Trigilia and Graziano Moramarco for their generous time, insightful discussions, and valuable advice, which greatly contributed to the development of this paper.

†Università della Svizzera Italiana, Faculty of Economics; email: *raffaella.intinghero@usi.ch*
‡Università della Svizzera Italiana & Swiss Finance Institute; email: *pietro.lazzaretto@usi.ch*
§Alma Mater Università di Bologna, Faculty of Economics; email: *paolo.manasse@unibo.it*

Non-Technical Summary

Why do people pay to publish an obituary when the death is already known and there are free alternatives ? This paper studies a side of obituaries that is often overlooked: in many cases, they are not only about honoring the dead, but also about promoting the living. In Italy, especially when a prominent figure passes away, newspapers fill up with condolence notices from individuals, businesses, and organizations. These paid messages can cost hundreds of euros and often serve no informational purpose—after all, the news is public. Instead, they act as public signals of closeness to the deceased, a way of associating one's own name or brand with a respected and admired figure.

This public display has much in common with other forms of status signaling. Just as wearing an expensive watch or sponsoring a prestigious event sends a message about one's social and economic position, an obituary alongside a famous death signals social proximity, influence, or prestige. The examples are striking: when industrialist Gianni Agnelli died, over 2,000 paid obituaries appeared in major Italian newspapers, ranging from prime ministers and luxury fashion houses to small local businesses eager to be seen in that space. Placement matters too—those whose messages appear next to the family's obituary gain the most visibility. By contrast, when a once-powerful politician dies in disgrace, the lack of obituaries itself becomes a statement.

To investigate this idea systematically, we gathered and analyzed roughly 144,000 obituary texts from 18 Italian newspapers, and compare publication patterns before and during the COVID-19 pandemic. The pandemic provided a unique natural experiment: mortality rates rose sharply in 2020, allowing us to observe how obituary behavior changed under extraordinary circumstances. If obituaries were mainly about spreading news of a death, the number per deceased should not have risen dramatically when deaths surged. Yet, we find that per-capita obituaries did increase significantly—driven almost entirely by notices for "popular" individuals, where the incentive to be publicly associated is strongest.

Our findings suggest that, in Italy at least, obituaries are more than private tributes; they are part of a marketplace for visibility, where economic motives and social incentives intersect. This reframing doesn't deny their emotional role—it adds another layer to it. By looking at an age-old tradition through an economic lens, we show that even acts of mourning can double as acts of marketing. And in doing so, we highlight how rituals of grief, much like other public behaviors, can be shaped by the desire to be seen, remembered, and connected to the symbols of prestige in a community.

I am not afraid of death. I just don't want to be there when it happens (Woody Allen, 1975)

1 Introduction

According to sociologists and moral philosophers, obituaries reveal "what counts as a value, virtue, or constituent of well-being for a particular type of person in a particular community" (Alfano et al., 2018). In this paper, we claim that obituaries also reveal something else: the more selfish behavior of economic calculus.

When friends, relatives, or simple acquaintances die, it is common, in Italy as in many other countries, to post a paid obituary. Depending on the place of residence and social status of the deceased, obituaries appear in national or local newspapers. Pages dedicated to obituaries are quite popular. For the elderly, getting to know if any acquaintance has died is often an important reason for purchasing the newspaper. It is also a major source of finance for the press. In a well-known sketch, Italian actor/writer Walter Valdi says *"in a newspaper I only read the obituaries page and that of movies. If someone I knew died, I go to his funeral. Otherwise, I go to the movies"*.

Obituaries posted by family members mainly serve the purpose of informing friends of the death of a relative. But this is not the only purpose of paid obituaries. When some prominent member of society dies, newspapers' pages fill up with "participations". These paid obituaries are typically posted by non-family members, who express condolences to the family. Their cost is quite hefty: depending on the text length and the type of newspaper, local or national, it ranges between 50 and 400 Euros. In these cases, the obituary does not convey information on the loss of a dear one, whose death is already public knowledge. This brings the question: why should a rational individual want to pay such a price? After all, if the aim is to express condolences to the family, a phone call, text message, or a post on social media would achieve the purpose at no cost. In this paper, we argue that obituaries often incorporate a "self-promotional" feature: they send the message that the person who posts the obituary, and whose name appears therein, was a "close acquaintance" of the deceased, particularly when the latter was a well-known (professionally or otherwise) individual. There is plenty of anecdotal evidence of this.

Journalist Guido Vigna, interviewed by the daily newspaper *Il Giornale*[1], recalls that the Italian (possibly world) record number of obituaries is probably held by Gianni Agnelli, a.k.a. L'Avvocato (The Lawyer), President of Fiat, President of Juventus Football Club, President of the Entrepreneurs' Association Confindustria, life Senator of the Republic, and Italian fashion icon. When he died on January 24, 2003, Vigna counted more than 2000 paying obituaries, appearing in all the most important national newspapers (*Corriere, Repubblica, Sole 24 Ore, Messaggero, Il Giornale*), not counting those published in foreign ones. These obituaries were posted by public and private entities, foundations, prime ministers and city mayors, sport and culture associations, newspapers' boards, trade unions, university departments, and so on. They also featured

[1]Stefano Lorenzetto, *Ha raccolto 100mila necrologi «Incredibile, nessuno muore»*, Il Giornale, February 8, 2015

some embarrassing obituaries, such as one reciting *"FIAT voluntas Gianni"*. And some unexpected ones: in a local newspaper, *La Provincia Pavese*, a barber-shop owner expressed *"his most vivid condolences to the Agnelli family"* so that the talk of the town presumably was: did he know the Avvocato personally?" (Manasse, 2003).

More recently, following the death of Silvio Berlusconi, the Italian Prime Minister, media tycoon and AC Milan President, on June 12, 2023, we counted 681 paying obituaries published only on the *Corriere della Sera*, featuring the gotha of media, sport, politics, finance, law firms. Interestingly, as for the previous case, *La Repubblica* also published on obituary from a rather obscure business from Sardinia: *"The owners of the XX Carpentry Shop, Gianni and Monica, and their collaborator Marco, deeply saddened, take part in the mourning for the loss of an extraordinary man, Dr Silvio Berlusconi"*.

Indeed, obituaries share a few characteristics with attending a funeral. First, what matters is to be seen, so that one's name (or brand) or its presence does not go unnoticed. Second, both activities aim at pretending close intimacy with the extinct. Italian novelist Giacomo Papi recalls on the newspaper *Il Foglio*[2] the case of Giulia Maria Crespi. Known as *La Tsarina* for her arrogant style, she was an entrepreneur, an aristocrat, a former owner of *Corriere della Sera*, the founder of the Italian Environment Fund (FAI), the owner of a noticeable art collection, including two huge Canaletto paintings. When she died on 19 July 2020, three pages filled up with 194 obituaries on the *Corriere*, for an estimated revenue of about 40 thousand euros. The position of the obituary, claims Papi, was very important for visibility: the closer to the family's obituary, the better. In the case of Donna Crespi, just next to the family's obituary appeared those of Urbano Cairo, current owner of the *Corriere*, followed by those of Ferrero (Nutella), Buitoni (Pasta maker), Armani, Zegna, Ferrè, Prada (Fashion), and so on.

In some cases, however, publicity can be bad, and this should be avoided. When disgraced Prime Minister and Socialist Party leader Bettino Craxi died exiled in Hammamet, Tunisia, on January 19, 2000, following the *Mani Pulite* corruption sentence, Craxi, once the most powerful and revered Italian politician in the 80's, received only a handful of obituaries: his family's, a trade-union secretary's and couple of comedians'. Indeed, Papi argues that *paid obituaries are the place where, at least in Italy, the geography of power is drawn. The map...that allows one to infer the networks of friendships, properties, shared interests, clienteles.*

This network/self-promotional feature of Italian obituaries seems to be somewhat different from the Anglo-Saxon tradition. There, newspapers publish longer articles of a completely different kind, whose purpose is to share a public memory. The *New York Times*' journalist Alden Whitman, considered the greatest columnist in the genre, took the habit of interviewing famous people in advance of their death, so as to register their opinion of their own life's main achievements. For this, a visit from him wasn't always welcome. And yet, the Anglo-Saxon tradition also provides examples of the "self-promotional" role of obituaries. Papi (2020) recalls a famous case of a tombstone in the Brooklyn Cemetery, inscribed with the following words: *"Here rest XX, loving father and*

[2]Giacomo Papi, *Nei necrologi si racconta il passaggio della storia, il mutare dei gusti e dei valori*, Il Foglio, July 26, 2020

husband, who was born in 1910 and died in 1974. His famous drugstore is still open, 24 hours a day".

Recent history shows that obituaries are also testimonials of great tragedies. On March 13, 2020, the Eco di Bergamo, a newspaper from a northern Italian town which was one of the earliest and worst affected by the coronavirus, published 10 pages of obituaries for the victims of the pandemic. On May 24, 2020, the New York Times remembered the milestone of 100,000 lives lost to Coronavirus in America by publishing the names of the dead and the memories of their lives, collected from obituaries across the US.

During the Pandemic that hit Italy in 2022, before spreading to other countries, mortality rates soared, and so did the number of per-deceased obituaries. This is shown in Figure 1. We ask: why should larger mortality rates be associated with the number of per-capita obituaries? We argue that the use of obituaries as a device for self-promotion accounts for this phenomenon. We exploit the dramatic natural experiment of the diffusion of the COVID-19 pandemics across Italian towns, in 2020, and analyze how this was reflected in the number and types of obituaries published in local and national newspapers. We show that the higher number of per-capita obituaries is entirely accounted for by the self-promotion motive.

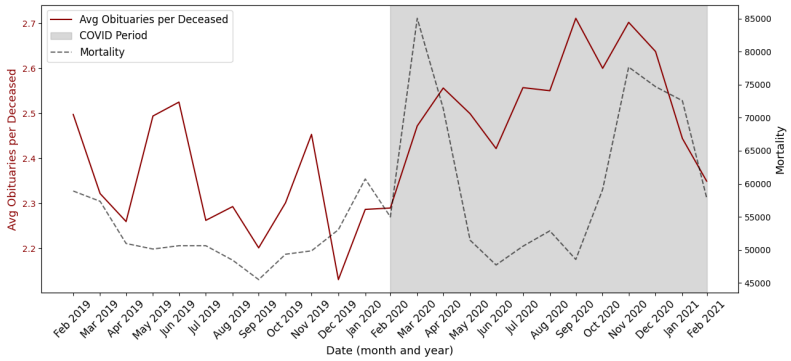

Figure 1: Average number of obituaries per deceased and monthly mortality

There is a large literature in sociology and moral philosophy investigating the lexicon of obituaries and how this reflects the evolving collective memory, the map of power and networks, cultural norms and stereotypes, and, in general, the evolving social values of the society. This literature dates back to the work of French sociologist Pierre Bourdieu (Bourdieu, 1988; Bourdieu, 1996) for whom obituaries summarize the cultural and moral values through which the elites celebrate themselves. Bourdieu (1996) examines the obituaries published between 1962 and 1965 in the *Annuaire de l'Association des Anciens Eleves de l'Ecole Normale Supérieure* to note how the different qualities praised and attributed to the deceased *Normaliens* reflect a precise hierarchy of social status based on academic success, social origin, and place of residence (Paris vs the province). More recently, Fowler & Bielsa (2007) explore how obituaries reflect societal values and hierarchies, focusing on the types of lives that are deemed noteworthy enough for pub-

lic commemoration in newspapers. By combining quantitative analysis with qualitative examination, they find a significant gender bias, with over-representation of men, with qualities that are often defined in terms of family roles rather than individual accomplishments. They conclude that obituaries reflect and reinforce social hierarchies and their values.

To our knowledge, our paper is the first economic analysis of obituaries' texts. Our aim is to investigate the self-promotional feature of obituaries in Italy.

The paper is also related to the idea of conspicuous consumption (Veblen, 1918; Corneo & Jeanne, 1997). Promotional obituaries have no intrinsic utility, as conspicuous consumption. However, rather than being a signal for an unobservable economic characteristic, such as income, they signal the *proximity* to a well-known deceased: this makes him/her a sort of "testimonial" for the name/brand of the signatory.

The paper proceeds as follows. In Section 2 we introduce a simple model describing the economic incentives to post (and to read) an obituary, and relate this to the "popularity" of the deceased. In Section 3 we describe the data. By scraping obituaries' websites, we collect the texts of all obituaries published in 18 local and national Italian newspapers in 2019, our pre-COVID period, and 2020, our post-COVID one, roughly 144,000 of them. Next, we discuss our empirical strategy. First, we use Natural Language Processing (NLP), text analysis, and classification algorithms in order to identify the keywords associated with the obituaries that are most likely to be motivated by advertising purposes and to identify the more likely "popular" deceased. Last, we exploit the "natural experiment" provided by the dramatic surge in deaths (and obituaries) due to COVID-19, and use regression analysis to test the main insight on the self-promotion model of obituaries. The Pandemic is indeed associated with a rise in per capita obituaries, and this is (almost) exclusively accounted for by "popular" deceased individuals. The final part of the paper checks the robustness of the results and concludes.

2 The Model

In the model (see Appendix A), we formalize the self-promotional incentive to publish and to read obituaries. Individuals differ in terms of their social popularity, which is a measure of how many people each one knows/is known by. In the "market for obituaries", there are two sides: publishers and readers. A reader spends time, at a cost, reading the newspaper obituary pages. She gets some utility if this activity is informative, which happens when she reads about someone she knew. This happens more frequently when the number of deaths rises, implying that more individuals will devote time to reading the obituary pages. Conversely, there are two types of publishers, family and non-family (participations). A family obituary simply informs readers of the death of a family member. A non-family type involves an economic calculus. Publishing an obituary comes at a cost, and provides an advertising /visibility benefit which is higher the more popular the deceased is and the larger the number of readers is. It follows that a sudden increase in the flow of deaths will be associated to a larger number of readers, increasing the benefit from promotional obituaries, and leading to more than proportional increase

in the number of obituaries, and to a larger per-capita number of obituaries for "popular" individuals.

3 The Data

Our dataset was obtained by scraping the website of *La Republica* dedicated to obituaries published in the editorial group's (GEDI) website[3]. We collected the texts of all obituaries published daily in 18 Italian newspapers in 2019, our pre-COVID period, and 2020-2021, our post-COVID one, roughly 144,000 entries. The data set covers two national newspapers (*La Repubblica, La Stampa*) and sixteen local newspaper (*Corriere delle Alpi, Gazzetta di Mantova, Gazzetta di Modena, Gazzetta di Reggio, Il Centro, Il Mattino di Padova, Il Messaggero Veneto, Il Piccolo, Il Secolo XIX, Il Tirreno, La Nuova Ferrara, La Nuova Sardegna, La Nuova Venezia, La Provincia Pavese, La Sentinella del Canavese, La Tribuna di Treviso*). Most newspapers distinguish between "obituaries", monthly and yearly "anniversaries" and "participations". The first are the ads typically published by relatives who announce the death of a family member and give information about the time, day, and place of funeral. "Participations" are typically posts that are attached to "obituaries", and express condolences to the family. From this text corpus, we have extracted information on the deceased (name, title, occupation, if available, place and date of birth and death), on the newspaper, date, and the place of publication, as well as the complete text. We have organized the data by the identity of the deceased, collecting together all the obituaries and participations that refer to each individual. This has required a massive work, as obituaries concerning the same person often appear on different newspapers, on different dates, sometimes close and sometimes far from each other. For example, anniversaries, remembrance ceremonies and so on may published years away from the day of death. Also, the same individual may appear under different names (nickname, titles, different order of name-surname) in obituaries, participations or anniversaries. We also have several cases of homonymy, and, to complicate matters further, we have many cases, particularly during the COVID springtime of 2020, where obituaries were published quite a long time after the day of death, because sometimes gatherings and funerals were not allowed, and only took place in the following months. All this has required double-checking with names and dates of birth and deaths in order to organize the data by the deceased's identity. In the end, we chose to consider only the obituaries published within a week of the day of death. Finally, we subjected all texts in our corpus to a time-intensive cleaning and pre-processing phase. This consisted of the removal of duplicates and uninformative tokens (e.g. dates, URLs, and nicknames). Then, we applied procedures commonly used in textual analysis, such as tokenization, stop-word removal, and stemming. Further details are provided in Appendix B.

In order to provide a first descriptive insight into both the temporal and geographical variability of obituary publication in Italy, we report in Table 1 the summary statistics of the number of obituaries per deceased.

[3] `https://necrologie.repubblica.it/`

Panel A shows that the average number of obituaries per deceased increased from 2019 to 2020, the year in which the COVID-19 pandemic spread across Italy.

Panel B shows substantial heterogeneity across newspapers. In some regions, in particular Sardinia, the number of obituaries and participations per deceased is much larger than elsewhere, for cultural reasons, as reflected in the high average reported by *La Nuova Sardegna*. Also, national newspapers tend to attract obituaries for more "famous" individuals, and accordingly show a higher number of obituaries and participations per deceased.

Table 1: Summary Statistics of Obituaries per Deceased

	Mean	Std. Dev.	Min	Max	Obs
Panel A: By Year					
2019	2.33	4.62	1	253	19988
2020	2.53	5.18	1	201	25061
2021	2.41	4.61	1	96	3991
Panel B: By Newspaper					
Corriere delle Alpi	1.12	2.24	1	95	1803
Gazzetta di Mantova	3.09	6.21	1	180	4209
Gazzetta di Modena	1.50	3.92	1	95	618
Gazzetta di Reggio	1.35	3.50	1	95	1423
Il Centro	1.05	1.55	1	95	3696
Il Mattino di Padova	1.23	2.57	1	95	1456
Il Messaggero Veneto	1.61	2.26	1	95	7540
Il Piccolo	2.20	3.20	1	95	4143
Il Secolo XIX	2.88	7.09	1	253	4262
Il Tirreno	1.43	3.36	1	95	1194
La Nuova Ferrara	1.18	2.61	1	95	1330
La Nuova Sardegna	4.53	6.59	1	115	8982
La Nuova Venezia	2.33	8.22	1	95	137
La Provincia Pavese	1.85	4.15	1	95	887
La Repubblica	2.33	4.39	1	95	2093
La Sentinella Del Canavese	2.17	3.19	1	15	41
La Stampa	2.70	7.13	1	201	5313
La Tribuna Di Treviso	2.21	7.16	1	95	184

Notes: "Mean", "Std. Dev.", "Min", and "Max" refer to the number of obituaries published per deceased. In Panel A, "Obs." refers to the number of deceased individuals associated with each year. By contrast, in Panel B, the unit of observation is the individual–newspaper pair. That is, individuals who received obituaries in multiple newspapers (240 cases) are counted once for each associated newspaper.

Figure 2 shows the highly unbalanced nature of our data, with more than 65.5% of individuals receiving only one obituary. The distribution of per-capita obituaries is strongly right-skewed, with a mean of 2.44, a median of 1, and a maximum of 253. In the empirical section we will take these features of the data into account.

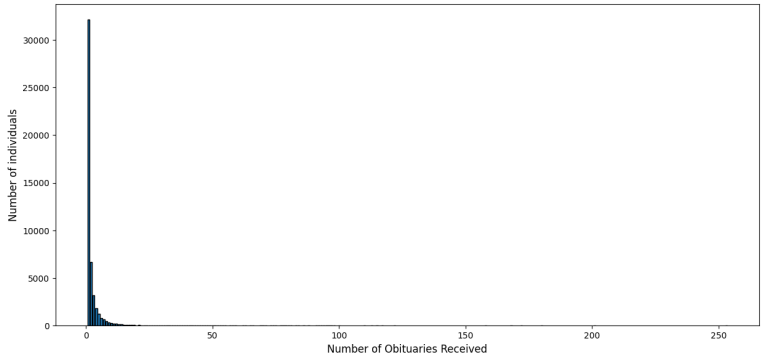

Figure 2: Distribution of individuals by number of obituaries received

4 Methodology and Empirical Analysis

The next section describes the methodology used to test the conjecture that the promotional motive in the use of obituaries explains the observation that during periods of high mortality the number of obituaries received by individuals rises, and this occurs for the most socially connected deceased. Unfortunately, to quote Sean Connery's Agent 007, we (do not) *only live twice*, so we cannot observe the counterfactual of how many obituaries a given individual would have received had she died in a different period. To make-up for this, we employ techniques for the analysis of texts. These allow us to classify *obituaries* into "family" and "non-family" types. After that, we classify *individuals* as more or less "popular", i.e. worth the cost of the obituary, depending on the number and the percentage of non-family obituaries received.

We proceed in three steps. The first step aims at identifying those words (stems) in the obituaries' text that flag a non-family as opposed to family type of obituary. To this end we need reduce the number of candidate stems dramatically. Because all obituaries share a large number of words related to condolences and parenthood, standard classification methods have little discriminating power. To address this, we start with the observation that non-family obituaries are generally published for individuals who receive many obituaries. Accordingly, we develop an original Bayesian algorithm that calculates, for each word, the probability that an individual receives any given number of obituaries, conditional on the word's appearance in the text. All stems are then ranked on the basis of this conditional probability for an increasing number of obituaries. Next, we employ an "optimal stopping rule" to determine the length of the keyword list to use as input in the classification exercise. At this stage we use a criterion for stem selection that applies to the entire data set and that does not require a manual labeling of observations.

The second step involves the manual labeling of a random sample of 5,062 obituaries, which we read and classify into the two classes of family vs. non-family types. On this sample, we train two supervised machine learning algorithms, restricting input features to our keyword list. We then use the trained models to predict classifications on the out-of-sample data, extending the labeling to the entire corpus of obituaries.

Since most individuals receive both types of obituary, we need the last step: moving from the classification of *obituaries* to that of *individual deceased*. We *define* as "popular" extincts, theoretically those for which the visibility gain justifies the publishing cost, as those individuals who receive a sufficiently large *share* of non-family obituaries. Finally, we use difference-in-difference econometric techniques to test the hypothesis that during the Pandemic these (and only these) individuals receive more per-capita obituaries than before the Pandemic. The next sections spell out the details.

4.1 A Bayesian Probability Algorithm

Our aim here is to classify the corpus of obituary texts into two categories: "family" and "non-family". To this purpose, we first identify a set of keywords that are most distinctive of non-family obituaries. The resulting list of keywords is later used as input for the classification exercise. Obituaries' texts present a special difficulty in this respect: words related to condolences, pain, family relationships are ubiquitous and appear almost always. To address this issue, we develop an original "Bayesian" algorithm that ranks all the words in the corpus, assigning greater weights to those terms who are more frequent for individuals receiving many obituaries. The assumption here is that individuals receiving a large number of obituaries are more likely to receive many non-family ones. This because the number of family members and family-type obituaries are homogeneous across individuals. This approach necessitates modeling the probabilities of observing a given number of obituaries per deceased, conditional on each stem appearing in the analyzed corpus. The idea is to assign a higher rank to stems that help predict a higher number of per capita obituaries, that is, for which the probability of observing n per capita obituaries, conditional on the word stem appearing in the text, increases with n.

Obituaries are typically brief texts comprising a limited number of words. This characteristic can result in significant variability in the word count across the corpus texts.

We construct "documents", indexed by $n = 1, 2, ..$ and denoted by d_n, by collecting together all the obituaries written for those individuals that receive exactly n obituaries. So, for example, d_2 is the collection of the texts published for individuals that receive 2 obituaries each. In our search for keywords, we are assuming that the texts pertaining to individuals who receive the same number of obituaries are "similar" and can be lumped together.

Define $p(s|n)$ as the probability of finding stem s in document d_n. I.e., the probability that the word appears conditional on an individual receiving n obituaries. We can apply Bayes' theorem to recover the probability that an individual receives n obituaries conditional on stem s appearing in the text, as $p(n|s)$:

$$p(n|s) = \frac{p(s|n) \cdot p(n)}{p(s)}, \tag{1}$$

where $p(n)$ is the ratio between the number of individuals receiving n obituaries and the total number of individuals, and $p(s)$ is the frequency of stem s in the corpus. In order

to address the problem of ubiquity of particular words, we modify these probabilities according to the measure of Term Frequency-Inverse Document Frequency (TF-IDF). This gives a larger weight to more rare stems. Formally, our operative conditional probability is

$$p'(n|s) = p(n|s) \log\left(\frac{|\mathcal{n}|}{\sum_{n=1}^{N} \mathbb{I}(p(n|s) \neq 0)}\right), \tag{2}$$

In this expression, the original probability is weighted by a term called the Inverse Document Frequency (IDF), where $\mathbb{I}(p(n|s) \neq 0)$ is an indicator function that takes the value 1 if $p(n|s) \neq 0$, and 0 otherwise. Consequently, the IDF term is the log of the ratio between the total number of documents in the corpus, in the numerator, and the number of documents in which the stem s appears, in the denominator. For example, for stems that appears in every document, irrespective of the number of obituaries associated, the ratio in the bracket is equal to one so that the weight is zero. Conversely, a stem which appears in only one document has a maximum weight equal to $\log(|\mathcal{n}|)$.

Ideally, in order to identify stems associated with non-family obituaries, we would like to pick those stems whose conditional probability rises with n. However, despite the previous adjustment, the corpus still contains a substantial proportion of rare stems, which unreasonably affect these probabilities and result in large non-monotonic "jumps" in the probabilities for close values of n.

To address this issue, we follow this approach. For each stem s, for all pairs[4] $n, n' > n$ we count the number of times when the stem's conditional probability of appearing in $d_{n'}$ exceeds that of appearing in d_n. That is, the occurrences of $p'(n'|s) > p'(n|s)$, for $n' > n$. The keywords that better identify non-family obituaries are those that have the highest count of positive variations, $NPV(s)$. This measure selects stems that exhibit consistent increases in conditional probability and ensures that the selected words are not rare in the corpus. The stems are then ranked according the number of positive changes, $NPV(s)$, see Appendix C.2.

Finally, a common problem in Text Analysis (Loughran & Mcdonald, 2011; Li et al., 2020; Shapiro et al., 2022 and Ash & Hansen, 2023) is that words can have different meanings in different contexts. For example, the stem "*partecip*" (to take part) appears both in non-family obituaries (to express condolences and sympathy towards the relatives of an important person) and in family obituaries (to convey gratitude to those who will participate in the funeral). Thus, *bi*-grams (pairs of successive terms) that capture a segment of the context, allow a better description of a corpus. We therefore repeat the analysis considering a bigram as our primary unit of analysis. Indeed, as will will show below, bigrams enhance the differentiation between family and non-family types. For example, bigrams that link stems such as *"partecip funeral", "partecip profond"*) help differentiate between family (expressing thanks for participating to the funeral) and non-family (expressing deep condolences) types.

Figures 3 and 4 show the *word clouds* of the most relevant stems and bigrams selected

[4]According to standard combinatorial calculus, all possible pairs of documents, such that $n' > n$, are given by $\binom{|\mathcal{n}|}{2}$. In our case, this yields 4,851 variations.

on the basis of their NPV measure. The size of each term (stem in both unigram and bigram forms, s) is proportional to the value of the associated NPV(s) function.

Figure 3 shows that stems with higher NPV values are predominantly associated with professional domains. Notable terms include "*professional*", "*avvoc*" (lawyer), "*sindacal*" (auditor), "*cda*" (board of directors), and "*dot*" (doctor).

Figure 3: Unigrams with the highest NPV values

The most significant bigrams are displayed in Figure 4, and align with the unigram findings.

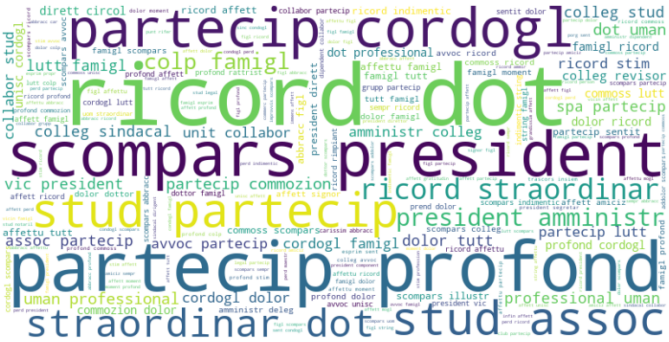

Figure 4: Bigrams with the highest NPV values

As with the unigrams, the most relevant bigrams are primarily related to professional fields. For instance, "*scompars president*" refers to the passing of a president, "*stud assoc*" denotes the office sharing the family's grief, "*stud partecip*" refers to the office taking part in the family's grief, and "*president amministrator*" indicates both the president and administrator expressing their condolences. A few bigrams, such as "*partecip profond*" (join the mourning) and "*partecip cordogl*" (take part in the grief), are expressions of condolence.

These results are consistent with the expectation that self-advertising obituaries

would emphasize professional ties while expressing condolences. The prevalence of professional terms underscores the importance of these themes in self-promotion within such obituaries.

4.2 Optimal List of Keywords

Having ranked *all* keywords in the texts, we now briefly discuss how we cut the unigram/bigram lists short and select an "optimal" number of keywords to use in the classification exercise in the next step. The advantage of this criterion is that it can be implemented for the entire sample without requiring any manual labeling of observations.

We start by assuming that if an individual receives *only one* ("single") obituary, then this *must* be a family type, so that ideally our list should *not* comprise keywords appearing in these obituaries. Based on this metric, we compare keywords appearing in "single" obituaries with those appearing for individuals receiving 2,3,..up to $n = 15$ obituaries (thereafter results stabilize). For each comparison, we evaluate lists containing the the first t_n NPV keywords. For example, the first comparison is between texts for individuals who receive 1 and those who receive $n = 2$ obituaries. Here we evaluate different lists of keywords containing the top $t = 5, ..3000$ terms in the list[5], and select the list length that score according di different criteria, call it t_2^*. Intuitively, these criteria specify that the "optimal" list should include the smallest possible number of keywords appearing in "single" obituaries, while including the largest number of those that appear in the more "numerous" one. We iterate this procedure comparing "single" and $n = 3$ individual and identify a new optimal list of t_3^* keywords, and so on up to to t_{15}^*. We run this procedure separately for bigrams and unigrams. By comparing the lists t_n^* for the different n based on a procedure described in Appendix D, we select two lists containing, respectively, the top NPV 275 bigrams and 85 unigrams stems.

4.3 Obituaries classification

The Bayesian probability algorithm and the "optimal stopping rule" help us reduce the overall vocabulary to a subset of keywords representing the most common bigram and unigram stems across texts of individuals who receive a large number of obituaries. Consequently, they should also be the most distinctive in identifying non-family obituaries.

In the present stage, we finally address the task of classifying obituaries into "family" and "non-family" types. This clearly requires the manual classification of a subset of the observations. To create the training set, we manually classified 5,062 obituaries into "family" and "non-family" categories. We classified as "family" all obituaries that announced a death or provided funeral details, such as the place and date. In contrast,

[5]For computational feasibility, instead of testing all possible list lengths, we evaluate performance over a predefined grid of list sizes. Specifically, we consider keyword lists of size 25, 50, 75, ..., up to 3000 for bigrams (in steps of 25), and 5, 10, 15, ..., up to 250 for unigrams (in steps of 5). The difference between these two intervals can be explained by the fact that bigrams tend to increase variability across texts, since it is more difficult for a given pair of words to appear in multiple documents compared to single words. In this context, bigrams require evaluation over longer lists of terms.

obituaries expressing condolences to the deceased's family, citing a company, institution, or firm, or highlighting a professional relationship between the writer and the deceased were classified as "non-family". To ensure impartiality in this classification task, we excluded from non-family obituaries those that express condolences but lack a formal signature with the signatory's full name. We assume that individuals who publish obituaries for promotional purposes carefully include their identifying information. In other words, the absence of a surname was used as an indicator of a lack of interest in being identified by anyone other than the deceased's family and as a sign of a close relationship with them.

To obtain a simple benchmark classification, we start using a "Naive" criterion: we classify an obituary as "non-family" if it contains at least one of the unigrams or bigrams appearing in the optimal list. Despite the high interpretability of this classification, this model has low power, as it may incorrectly classify many family obituaries as non-family. For example, our keyword "professor", which appears in our optimal list, could also be present in family obituaries, in expressions such as "the family announces the death of Professor "name" ". In this case, the Naive algorithm, not taking account of the context, would wrongly classify this obituary as non-family.

To overcome this limitation, we train two machine learning models, Random Forest and Support Vector Machine, on a subset of manually labeled obituaries. Both algorithms use our keywords as "features"— that is, variables on which the classification is based. This allows them to consider combinations of possibly many keywords (e.g., "professor" and "partecip cordogl," meaning "take part" in the grief) to classify obituaries as non-family or family, rather than looking at each keyword individually.

For both Random Forest and Support Vector Machine, we restricted the vocabulary to the optimal lists of unigrams and bigrams selected in the previous step. This helped reduce feature complexity and noise, lowering the risk of overfitting — a situation where a model lacks out-of-sample generality because it is too closely fitted to the training data.

The performance metrics of the Random Forest and Support Vector Machine are presented, respectively, in Table 10 and 11 in Appendix E.2, where we also show the most relevant keywords for Random Forest classification (Figure 6). According to intuition, the most important bigrams and unigrams for identifying non-family obituaries refer to corporate roles and duties. Additional technical details about the models' configuration are provided in Appendix E.1.

Overall, both classifiers show high performance in the out-of-sample prediction, with accuracy, the percentage of correctly classified observations, reaching 90% for the Random Forest and 87% for the Support Vector Machine. This means that, when we employ the two algorithms to predict labels for a 100-randomly-selected subset of obituaries from our manually labeled sample, they correctly predict classifications for 90 and 87 of these.

We employ each trained model separately to classify all obituaries in our dataset, obtaining two distinct classifications in addition to the one based solely on keyword counts (the Naive algorithm).

Finally, the classification of obituaries into "non-family" and "family" types allows us to identify the "well-connected" or *popular* deceased, those that are worth spending

money by self-interested survivors for promotional purposes. We define them as those deceased who satisfy two criteria: they receive more than one obituary, and at least 50 percent of the obituaries received are non-family type (we experiment with different thresholds). The three classifiers, Naive, Random Forest, and Support Vector Machine, roughly identify the same individuals as "popular". Specifically, Table 12 in Appendix E.2 shows that, using the Naive classifier as a benchmark, 94% of individuals are classified in the same way when moving from the Naive to the Support Vector Machine classification, and 90% when moving from the Naive to the Random Forest. Furthermore, what makes these classifications convincing is that, when moving from the Naive to one of the two Machine Learning classifiers, the latter tend to be more restrictive in classifying individuals as 'popular'. In fact, no individual classified as "non-popular" by Naive is classified as 'popular' by either RF or SVM. While 4769 (2564) individuals that Naive classifies as 'popular' are instead labeled "non-popular" by RF (SVM).

Table 2 shows the number of individuals classified as "popular" under each threshold across the three classifiers.

Table 2: Number of *popular* individuals across classifiers and thresholds

Threshold	**Random Forest**	**SVM**	**Naive**
$> 50\%$	4,928	7,133	9,697
$> 55\%$	4,799	7,055	9,649
$> 60\%$	4,053	6,359	9,068

Notes: Popular individuals are defined as those for whom the percentage of promotional obituaries received exceeds the indicated threshold.

To give an idea of how many obituaries popular individuals tend to receive, Table 3 shows that, as we look at groups receiving a higher number of obituaries, the share of individuals classified as "popular" increases .

Table 3: Share of *Popular* individuals (%) among those receiving more than k obituaries

Number of Obituaries ($>k$)	**Obs.**	**Random Forest**	**SVM**	**Naive**
> 1	16902	29.2	42.2	57.4
> 2	10239	40.8	57.4	72.1
> 3	7073	43.1	60.5	74.1
> 4	5236	49.6	67.7	80.4
> 5	4020	51.6	70.6	82.3
> 10	1539	64.0	81.1	88.4
> 15	783	68.5	85.1	90.7
> 25	314	74.8	86.6	91.1

Notes: Popular individuals are those for whom more than 50% of their obituaries are classified as promotional. "Obs." indicates the number of individuals receiving more than k obituaries.

Although this table uses the 50% threshold of non-family vs family obituaries to

define a popular deceased, the pattern holds across other thresholds as well.

5 Regression analysis

In this section, we test the hypothesis of self-promotional behavior in obituaries. We employ a Difference-in-Differences (DiD) strategy. Our conjecture is that periods of high mortality are associated with increased interest in death-related content, a higher number of readers of the newspapers' obituary sections, and, consequently, a higher incentive to publish obituaries for the deceased whose public visibility attracts more attention.

To empirically validate this mechanism, we exploit the increase in mortality caused by the COVID-19 pandemic. Specifically, we assess whether this event was associated with a rise in the number of per-capita obituaries, and whether this increase is driven by "popular" individuals.

In our Difference-in-Differences design, the treatment group consists of "popular" individuals identified by text analysis, while the control group includes all the other individuals.

The estimated model can be formalized as follows:

$$\log(Y_{i,t}) = \alpha_0 + \delta P_{i,t} + \eta Pop_i + \beta(P_{i,t} \times Pop_i) + \theta N_i + \gamma T_t + u_{i,t} \tag{3}$$

The dependent variable $Y_{i,t}$ is the total number of obituaries received by individual i, who died on day[6] t. We apply a logarithmic transformation to smooth the skewed distribution typical of count variables. The explanatory variables are as follows: a dummy variable (*Pandemic*, $P_{i,t}$), which takes the value of one if the death of individual i occurred after the onset of the COVID-19 pandemic in Italy[7], capturing the post-treatment effect; a time-invariant dummy variable (*Popular*, Pop_i), which identifies the treatment and control groups, taking the value of one if individual i is classified as "popular" based on the methodology described previously; and the interaction term (*Pandemic* × *Popular*, $P_{i,t} \times Pop_i$), whose coefficient captures the differential effect of the pandemic on the number of obituaries for popular individuals relative to their counterparts. We also include fixed effects for the newspapers in which individual i's obituaries were published (N_i), and for the month of publication (T_t), in order to control, respectively, for systematic differences in newspaper type and coverage, and for seasonal effects.

Our conjecture implies that the estimated coefficient of the treatment variable (β) is significantly different from zero and positive, while the coefficient of the Pandemic, $P_{i,t}$ (δ) should not be significantly different from zero

In all the specifications, individuals are identified as "popular" if they receive more than one obituary and have a share of non-family obituaries exceeding 50%. These individuals are our treatment group, while the control group comprises all the others. This definition addresses the unbalanced nature of our data — where 65.5% of individuals

[6]Since the date of death is not available for all individuals, we use the date of the first obituary received for each individual as a proxy.

[7]We assume February 1, 2020 as the start date of the pandemic.

receive only one obituary[8]. Table 4) below presents the results of our estimates, based on obituary classifications obtained using the three different classifiers described in the previous section (Naive, Random Forest, and Support Vector Machine).

Table 4: Pandemic Effects on Per Capita Number of Obituaries

	(1) **Naive Model**	(2) **Naive Model**	(3) **RF Model**	(4) **RF Model**	(5) **SVM Model**	(6) **SVM Model**
Pandemic	0.0061	0.0077*	0.0119**	0.0137**	0.0052	0.0064
	(0.0043)	(0.0043)	(0.0052)	(0.0053)	(0.0047)	(0.0048)
Pop	1.1915***	1.1914***	1.1648***	1.1647***	1.2173***	1.2171***
	(0.0118)	(0.0118)	(0.0179)	(0.0179)	(0.0143)	(0.0143)
Pandemic×Pop	0.0348**	0.0348**	0.0598***	0.0599***	0.0476**	0.0477**
	(0.0155)	(0.0155)	(0.0232)	(0.0232)	(0.0186)	(0.0186)
_cons	-0.1867***	-0.1798***	-0.2384**	-0.2358**	-0.1983***	-0.1922***
	(0.0412)	(0.0420)	(0.0959)	(0.0959)	(0.0638)	(0.0642)
Newspaper Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month Fixed Effects	No	Yes	No	Yes	No	Yes
R-squared	0.551	0.551	0.401	0.401	0.494	0.494
Observations	49040	49040	49040	49040	49040	49040

Notes: Standard errors, reported in parentheses, are robust. *p<0.10, **p<0.05, ***p<0.01.
Pop is defined as individuals for whom more than **50%** of their obituaries are classified as promotional according to each classification model.

The table presents six columns, reporting the estimates for the three classifications (Naive, Random Forest, Support Vector Machine), each including/excluding a month-fixed effect. The results are consistent with our conjecture. First, the coefficient of the variable *Pop* (η) is positive and statistically significant at the 1% level across all specifications, confirming that individuals in the treatment and control groups differ structurally in terms of the average number of per capita obituaries. In other words, individuals classified as "popular" on the basis of our text analysis indeed receive a larger number of obituaries, about 3.3 times higher than the others (=exp(1.2)). Second, the coefficient of Pandemic δ, capturing the post-treatment effect, with the exception of the RF specification with a monthly dummy, is either weakly significant or not significant and always very small. Conversely, the variable of interest, the coefficient capturing the treatment effect β, is consistently positive, quantitatively large, and significant (at least at the 5% level) through all specifications. According to these estimates, the increase in per-capita obituaries for the popular individuals, relative to the others, during pandemic ranges between 3.56 percent ($= exp(0.035) - 1) * 100$) and 6.18 percent ($= exp(0.06) - 1) * 100$).

Interestingly, the estimated magnitude of the treatment effect β is higher for those classifiers (RF and SVM) that are more restrictive in classifying individuals as popular. This is because the least restrictive Naive model, based on keyword presence, is more prone to incorrectly classifying individuals as popular (false positives), whereas machine learning models may be more prone to false negatives (incorrectly classifying individuals as non-popular).

[8]This relies on the assumption that each individual receives at least one "family" obituary.

6 Robustness

In this section, we want to rule out the possibility that our results are driven by some alternative mechanisms related to the COVID-19 pandemic. For example, attending a funeral for someone "popular" may become dangerous or prohibited, leading to a larger use of obituaries. Alternatively, the Pandemic could hit proportionally less the richer and more connected individuals, leading to a concentration of obituaries. Although these mechanisms are coherent with our conjecture of promotional incentive, we want to make sure that our findings do not depend on the Pandemic-based specification of the "treatment". Specifically, we test whether, in general, periods characterized by a recent rise in the overall number of obituaries published daily are associated with an increase in the number of *per capita* obituaries, and whether this effect is larger for popular individuals. The idea is that self-promotional behavior is more likely when the "obituary market" is larger — that is, when the number of published obituaries and of potential readers has risen recently, so that advertising one's connection to the popular deceased becomes more profitable. To capture this effect, we compute a variable, termed *Market Size*, defined as the average daily number of obituaries[9] published across the newspapers in our dataset during the month preceding the first day on which an individual receives an obituary[10]. Similar results hold considering longer periods before the first obituary.

Table 5: Market Size Effects on Per Capita Number of Obituaries

	(1) **Naive** Model	(2) **Naive Model**	(3) **RF** Model	(4) **RF Model**	(5) **SVM** Model	(6) **SVM Model**
Log(Market Size)	0.0164	0.0318***	0.0200*	0.0416***	0.0153	0.0329**
	(0.0100)	(0.0121)	(0.0119)	(0.0142)	(0.0108)	(0.0130)
Pop	1.0435***	1.0462***	0.5593**	0.5580**	0.8582***	0.8619***
	(0.1709)	(0.1713)	(0.2621)	(0.2632)	(0.1995)	(0.2001)
Log(Market Size)×Pop	0.0325	0.0320	0.1241**	0.1243**	0.0749*	0.0742*
	(0.0331)	(0.0332)	(0.0509)	(0.0511)	(0.0387)	(0.0388)
_cons	-0.2680***	-0.3363***	-0.3353***	-0.4378***	-0.2749***	-0.3545***
	(0.0663)	(0.0746)	(0.1151)	(0.1215)	(0.0856)	(0.0929)
Newspaper Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month Fixed Effects	No	Yes	No	Yes	No	Yes
R-squared	0.551	0.551	0.401	0.401	0.494	0.494
Observations	49040	49040	49040	49040	49040	49040

Notes: Standard errors in parentheses are robust. *p<0.10, **p<0.05, ***p<0.01.
Pop is defined as individuals for whom more than **50%** of their obituaries are classified as promotional according to each classification model.
Market Size is defined as the average daily number of obituaries published in the considered newspapers during the month preceding the date on which each individual received their first obituary

For four out of six specifications in Table 5, the estimated elasticity of the per capita

[9]The average was calculated on the basis of a moving average that considers only the days actually available in the dataset, avoiding biases caused by missing data. For the first day of observation (February 20, 2019), the variable was set equal to the number of obituaries published on that day.

[10]This period is approximated as the 30 days preceding the event. We provide a robustness check for this definition in the following section.

number of obituaries with respect to the average "obituary market size" is positive and statistically significant, suggesting individuals who die following a period when many obituaries are published tend to receive a larger number of per capita obituaries. The coefficient for the variable *Pop*, which denotes individuals defined as popular, is positive and significant at least at the 5% level across all specifications, confirming the findings obtained previously. The crucial estimate of the interaction term between *Log(Market Size)* and the *Pop* dummy is positive and statistically significant in both the Random Forest and SVM models, and positive but not statistically significant in the Naive model. As before, when the more restrictive classifiers are used, the effect of market size on the number of per capita obituaries for famous individuals is stronger and statistically more significant compared to less restrictive classifiers. Overall, while these results are less stable than those obtained earlier, they confirm the presence of a self-promotional motive in obituaries. The next two tables (Tables 6 and 7) present other robustness checks related to more stringent definitions of a "popular" individual, requiring that the share of non-family obituaries to rise to 55 and 60 percent, respectively. The effect is that now the interaction terms is significant also in the Naive model, the classification upward bias for "popular" being reduced. The effect on the more "parsimonious" classifiers, RF and SVM, is to reduce the magnitude of the estimated coefficient, as the number of deceased classified as popular falls.

Table 6: Pandemic Effects on Per Capita Number of Obituaries

	(1) **Naive Model**	(2) **Naive Model**	(3) **RF Model**	(4) **RF Model**	(5) **SVM Model**	(6) **SVM Model**
Pandemic	0.0063	0.0079*	0.0105**	0.0122**	0.0047	0.0060
	(0.0043)	(0.0044)	(0.0053)	(0.0054)	(0.0048)	(0.0049)
Pop	1.1798***	1.1797***	1.1143***	1.1143***	1.1950***	1.1948***
	(0.0118)	(0.0118)	(0.0177)	(0.0177)	(0.0143)	(0.0143)
Pandemic×Pop	0.0345**	0.0345**	0.0735***	0.0735***	0.0512***	0.0513***
	(0.0155)	(0.0155)	(0.0231)	(0.0231)	(0.0186)	(0.0186)
_cons	-0.1901***	-0.1842***	-0.2552**	-0.2524**	-0.2085***	-0.2025***
	(0.0430)	(0.0437)	(0.1043)	(0.1043)	(0.0694)	(0.0698)
Newspaper Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month Fixed Effects	No	Yes	No	Yes	No	Yes
R-squared	0.543	0.543	0.382	0.383	0.482	0.482
Observations	49040	49040	49040	49040	49040	49040

Notes: Standard errors, reported in parentheses, are robust. *p<0.10, **p<0.05, ***p<0.01.
Pop is defined as individuals for whom more than **55%** of their obituaries are classified as promotional according to each classification model.

Table 7: Pandemic Effects on Per Capita Number of Obituaries

	(1) **Naive Model**	(2) **Naive Model**	(3) **RF Model**	(4) **RF Model**	(5) **SVM Model**	(6) **SVM Model**
Pandemic	0.0065	0.0084*	0.0145***	0.0164***	0.0047	0.0064
	(0.0047)	(0.0048)	(0.0056)	(0.0057)	(0.0051)	(0.0052)
Pop	1.1144***	1.1144***	0.9948***	0.9948***	1.1080***	1.1078***
	(0.0124)	(0.0124)	(0.0195)	(0.0195)	(0.0155)	(0.0155)
Pandemic×Pop	0.0367**	0.0366**	0.0695***	0.0695***	0.0592***	0.0595***
	(0.0162)	(0.0162)	(0.0256)	(0.0256)	(0.0201)	(0.0201)
_cons	-0.1947***	-0.1887***	-0.2720**	-0.2713**	-0.2226***	-0.2152***
	(0.0495)	(0.0502)	(0.1147)	(0.1146)	(0.0803)	(0.0806)
Newspaper Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Month Fixed Effects	No	Yes	No	Yes	No	Yes
R-squared	0.495	0.496	0.325	0.325	0.427	0.427
Observations	49040	49040	49040	49040	49040	49040

Notes: Standard errors, reported in parentheses, are robust. *p<0.10, **p<0.05, ***p<0.01.
Pop is defined as individuals for whom more than **60%** of their obituaries are classified as promotional according to each classification model.

In order to validate the treatment identification strategy underlying our Difference-in-Differences specifications, we conducted a placebo exercise. This test verifies that, during the pre-treatment period, there were no significant differences in the outcome variable between popular and non-popular individuals. To implement this, we restricted the data to the pre-pandemic period (February 20, 2019 to January 31, 2020) and re-estimated the baseline regression models, assuming a fictitious treatment occurred at the midpoint of this period. We constructed a binary variable (*Placebo*) equal to 1 for individuals who died after August 11, 2019 (the midpoint of the pre-pandemic period), and 0 otherwise. The fictitious treatment is then represented by the interaction term ($Placebo \times Pop$). We conducted this exercise for all three classification methods (Naive, RF, and SVM) and for the three thresholds used to define "popular" individuals (50%, 55%, and 60% of non-family obituaries). In all specifications, the coefficient of the interaction term is negative and not statistically significant in 8 out of 9 specifications. This finding supports the assumption that the treatment and control groups did not exhibit different trends in the outcome variable prior to treatment, which would have otherwise indicated a violation of the parallel trends hypothesis. These results are available upon request.

7 Conclusions

This paper studies the use of paid obituaries for for self-promotional purposes through the lenses of text analysis. We look the case of Italy, a country were net-work are pervasive, during the Covid Pandemic. The mere existence of a market for paid obituaries is difficult to rationalize in a world of free communication. And so is the observation that per-capita obituaries increased considerably during the Pandemic, mostly related to texts containing references to economic activities and roles. We bring anecdotal evidence together with the text analysis and D-i-D regression. We identify the keywords signaling an economic motive behind the obituaries, and through them identify the deceased individuals who

better serve the purpose. We show that in Italy economic incentives indeed permeate the publication of these texts: advertising one's proximity with a well-known deceased explains the rise in pro-capita obituaries in the post Pandemic period.

Appendix A: The Model

We sketch the simplest model that delivers the message. Population is constant and normalized to 1. There is a constant flow of deaths (and births) per unit of time , D. There is a continuum of individuals who differ according to their social "popularity" v, and their economic conditions c. For simplicity these traits are assumed to be independently distributed in the population, although nothing hinges upon this assumption. Popularity measures how well-connected is the individual: an agent's v measures the mass of people that knows him, either personally or by fame. We also assume that well-known individuals carry a good reputation, so that being associated to them is "good business" because their network of acquaintances is large (they are good "testimonials"). We ignore cases of well-known individuals, such as disgraced politicians, that everyone wants to shy away from. Economic condition is measured by the (opportunity) cost of spending money on an obituary c, which is higher the larger share of this expense on the individual income. In short, high v individuals are "popular", while high c ones are "poor". Each deceased has one relative. We assume all relatives post one obituary for the family's deceased, irrespective of her characteristics v. This is done in order to inform friends and acquaintances, and there is no self-advertising motive in this decision. Thus,in each period there will be D distinct obituaries per unit of time from family members. Each individual is characterized by his social popularity and economic condition, *i(v,c)* , and is located on a circle of circumference 1. He is known by *v/2* people on his right and *-v/2* people on his left, for a total mass *v*. Social popularity v is distributed with cumulative distribution *F(v)*, which we assume to be independent of characteristic c. Deaths occur randomly, so when agent *i(v,c)* dies, the probability that a survivor *x(v,c)* knows him, i.e. that she is located within his range of acquaintances, is *F(i+v/2)-F(i-v/2)* which, for a uniform distribution on [0,1] is equal to *v*.

A.1 Readers of Obituaries

We assume that the typical newspaper reader *x(v,c)* simply chooses whether or not to devote a fixed amount of time to reading the obituary page. If he does, he goes through the ads and finds out who died. When one obituary on deceased i is published, the reader has probability v that this will concern someone she knew, in which case, reading the obituary provides useful information and utility $\bar{u}$. If the deceased is unknown to the reader, reading the obituary provides no information and zero utility. Devoting time to reading the obituary page comes at the fixed opportunity (time) cost τ. When D obituaries on $different$ deceased are published, the expected utility from reading the obituary page is $EU = Dv\bar{u} - \tau$. Letting

$$\delta = \tau/\bar{u} < D \tag{4}$$

denote the ratio between the opportunity time cost and the benefit from reading the obituary page, an individual will do so when $v > \delta/D$. Thus, the total number of readers per unit of time will be

$$Readers = Prob(v > \delta/D) = 1 - F(\delta/D) = 1 - \delta/D \tag{5}$$

where we assume that F is uniform in the interval [0,1]. Note that as the number of family obituaries (deaths) rises, the number of readers grows.

A.2 Posters of Obituaries

For a non-family member, posting an obituary (a "participation" to the family obituary, p) for the deceased v is based on an advertising calculation. The participation's signatory derives a benefit that increases with the number of readers as more people are informed of the signatory's proximity to the deceased. For example, every reader generates a benefit $b(v)$, $b'(v) > 0, b(0) = 0, b(1) = 1$, say $b(v) = v$. The opportunity cost of an obituary c is distributed with c.d.f. *G(c)*. The profit that accrues to an individual of type-c from posting a participation for deceased-v is therefore

$$\pi(c, v; D) = b(v)Readers - c = b(v)(1 - \delta/D) - c \tag{6}$$

Hence the number of individuals willing to post a participation for v will be

$$p(v, D) = Prob(c \leq b(v)(1 - \delta/D)) = G(b(v)(1 - \delta/D)) = b(v)(1 - \delta/D) \tag{7}$$

again assuming that *G(c)* is uniform in [0,1]. More famous (high v) people will get more participations. If we assume that the benefit function is linear $b(v) = v$, the total number of participations is

$$P = (1 - \delta/D) \int_0^1 v \, dv = \frac{(1 - \delta/D)}{2}$$

The total number of obituaries, including those of family members, is therefore O $=P+D$

$$O(D) = \frac{1}{2}(1 - \frac{\delta}{D}) + D \tag{8}$$

Assuming that the relative opportunity cost of reading the obituary page is small enough, $\delta < D$, the total number of obituaries rises more than proportionally with the number of deaths. This because, in addition to family obituaries, participations rise as their benefits increase due to a larger mass of readers. This effect should be entirely due to the rise of participations for "more popular" (high-v) deceased.

Appendix B: Cleaning work and Pre-processing

In order to use the data for text analysis we have made a number of interventions. The first, referred to as *cleaning work*, addresses specific issues related to the text in the analyzed corpus, while the second, *processing*, follows more standard procedures common to textual analyses.

B.1 Cleaning Work

To start, we excluded from the analysis the obituaries published long after the death (one month and one year), in order to recall the memory of the extinct (these are called "notices of the thirtieth day" and "anniversaries", respectively) . These make 12.9% of all the obituaries in the dataset.

A key task is the unique identification of each individual via an identification number. Variations in the reporting of names poses many challenges, such as differences in the order of first and last names or the inclusion of titles (e.g., "*Dr.*" meaning "doctor," "*Ved.*" meaning "widow"). A systematic cleaning process was employed to resolve these issues: names were alphabetically ordered and capitalized, while professional titles and marital status abbreviations were removed to ensure consistency.

We addressed homonymy cases by associating names with their respective provinces of residence. This information is always available, unlike other potentially useful variables, such as birth or death dates. The assignment of unique identification numbers based on names and provinces effectively solves the homonymy issue, allowing to distinguish different individuals sharing the same name and surname.

Subsequent steps involved the identification and removal of duplicate obituaries. A preliminary check revealed 57 duplicate entries, which have identical names, obituary texts, and identical values across all other variables. Further analysis uncovered a regional duplication issue: obituaries of individuals from the Sardinia region were erroneously listed under Sicily as well. Manual verification confirmed that Sardinia was the correct region of residence, leading to the removal of 97 additional duplicates.

The next cleaning step eliminated uninformative text elements such as names, dates, geographical references and website URLs commonly included by funeral service providers. The elimination of these elements reduces noise in the text and facilitates the identification of keywords pertinent to non family obituaries. In eliminating these names we were careful not to remove terms with ambiguous meanings (e.g., *Cuoco*, which can mean "chef" or serve as a surname). To this end, we used a comprehensive lists of Italian first names and surnames[11], which we manually verified.

Obituaries often contain names in the form of nicknames, diminutives, or typographical errors. In order to identify and remove such terms, we calculated the Cosine Similarity measure between each word of the corpus and the list of Italian proper names available on GitHub. This generated a list of candidate nicknames that was then manually reviewed to ensure no relevant terms were inadvertently removed.

[11]The two lists used are available at `https://github.com/napolux/paroleitaliane`, comprising 9,000 first names and 38,000 common surnames in Italy.

B.2 Pre-processing

The pre-processing phase involves the application of procedures commonly used in textual analysis: tokenization, stop-word removal and stemming. First, tokenization transforms the data into lists of individual words to facilitate systematic analysis. The removal of stop words eliminates frequently occurring terms that convey minimal informational value, such as definite and indefinite articles, personal pronouns, and conjunctions. Custom rules were applied to address specific cases, for example transforming "*s.p.a.*", denoting "joint-stock company", into *spa.* to preserve the meaning. Incorrectly spelled accented vowels were corrected to their proper forms to avoid distorting word frequency or occurrence counts due to typographical errors (e.g., "*professionalita'*" was replaced with "*professionalità*").

For the same reason, a list of 33 common professional abbreviations was standardized by replacing them with their full forms (e.g., "*Ing.*" was replaced with "*Ingegnere*" meaning "engineer"; "*Prof.*" with "*Professore*" for "professor"; and "*Rag.*" with "*Ragioniere*" for "accountant"). Additionally, all numerical characters, special characters, and words containing both numerical and textual superscripts were removed (e.g., "*la classe* $5^{\underline{a}}A$", referring to a school classroom, resulted in the removal of "$5^{\underline{a}}$A").

The final step in textual pre-processing involved a decision between lemmatization and stemming. Both techniques aim to standardize word forms by reducing inflected or derived words to their base form, ensuring that word frequency counts are not affected by morphological variation. Stemming reduces words to their root by eliminating suffixes, whereas lemmatization is a more sophisticated approach, reducing words to their lemma while considering grammatical context.

In this application, stemming was preferred over lemmatization due to its superior ability to handle issues arising from the presence of typographical errors and regional dialectical expressions. Consequently, stemming was applied to the entire corpus, with words like "*professore/professoressa*" reduced to "*professor*" and "*adorato/a*" reduced to "*ador*" ("loved one").

Appendix C: Text Analysis Framework

C.1 Definition of a document

Formally, let $\mathbb{N}$ denote the set of natural numbers, and let $\mathcal{n}$ be the set of containing all the observed numbers of per-capita obituaries n, which range from zero to a maximum of N so that :

$$\mathcal{n} = \{n \in \mathbb{N} \mid 0 < n \leq N\}, \quad |\mathcal{n}| = 99 \tag{9}$$

In our sample the most "popular" individual receives $N = 253$ obituaries. The "cardinality" of $\mathcal{n}$ (denoted $|\mathcal{n}|$) is the number of distinct values taken by n, which equals 99 in the dataset.
A document d_n is thereafter defined as the concatenation of all obituaries received by all the individuals which receives n obituaries, for $n \in \mathcal{n}$. As a result, the number of documents equals to the the number of different values taken by per-capita obituaries, $|\mathcal{n}|$. According to *Zipf's law* [12] a corpus typically contains only a small number of high-frequency words and many low-frequency words. This is corroborated by our corpus of obituaries, where over 62% of stems appear in only one obituary. To reduce the sparsity in the document-term matrix, we consider only stems appearing in more than one obituary. Consequently, we remove 22,918 stems from the corpus, which originally contained 36,891 unique stems. The resulting vocabulary, V, consists of 13,973 stems, s.

$$V = \{s \in \mathbb{N} \mid 0 < s \leq S\} \tag{10}$$

where $S = 13,973$.

C.2 Definition of the Number of Positive Variations

Formally, define the "Number of Positive Variations" for a generic stem $s \in V$ as:

$$\text{NPV}(s) = \sum_{(m,j)\in\mathcal{B}} \mathbb{I}\left(p'(n = m|s) - p'(n = j|s) > 0\right), \tag{11}$$

where:

$$\mathbb{I}\left(p'(n = m|s) - p'(n = j|s) > 0\right) = \begin{cases} 1 & \text{if } p'(n = m|s) > p'(n = j|s), \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Thus, a stem's NPV is the count of times when the stem's conditional probability of appearing in a document with a larger number of per- capita obituaries exceeds the stem's conditional probability of appearing in a document with a lower number of per-capita obituaries, for all possible pairs of documents. The stems are then ranked according to their NPV.

[12] Described in Loughran & Mcdonald (2011)

Appendix D: Keywords and Threshold Selection

In this appendix, we describe how we determined the optimal number of bigrams and unigrams employed both as keywords for the Naive algorithm and as input features for the Random Forest and Support Vector Machine algorithms.

In this context, the presence or absence of the keywords in an individual's obituary is treated as a binary decision test.

We proceed as follows. Let i denote an individual appearing in the observed sample of obituaries, where i ranges from 1 to K, the total number of individuals, which is 49,040. I denotes the set of all individuals in the sample:

$$I = \{i \in \mathbb{N} \mid 0 < i \leq K\} \tag{13}$$

where $\mathbb{N}$ is the subset of natural numbers composed of the identification codes of i.

Let n_i denote the number of obituaries received by individual i, where $n_i \in \mathbb{N}$. If an individual receives exactly one obituary, he is classified as 'Type 1' (denoted $\mathrm{T_i^I} = 1 \iff n_i = 1$), while if he receives at least $n > 1$ obituaries he is classified as 'Type n' (denoted $\mathrm{T_i^I} = n \iff n_i \geq n$). The population is therefore iteratively categorized into two distinct groups based on the binary variable $\mathrm{T_i^I} \in \{1, n\}$, where $n = 2, 3, 4...$

The idea is the the list of keywords of optimal length, based on NPV rank, should contain the lowest number of keywords that appear for deceased who receive only one obituary (Type 1) and the highest for those who receive $n > 1$ We start with analyzing the simplest case where $n = 2$. For a given list of the first t terms in our NPV rank, we look into the obituaries received by each individual i, and check whether at least one of the keywords appears therein or not. If it does, the result of the check is positive $\mathrm{R_i} = R^+$, while if it does not, the result is negative $\mathrm{R_i} = R^-$. It is therefore natural to define as False Positives Fraction (FPF), the percentage of Type 1 individuals whose obituaries contain at least one of our ranked keywords:

$$\mathrm{FPF_{t,1}} = P(\mathrm{R_i} = R^+ | \mathrm{T}_i^I = 1) \tag{14}$$

Similarly the True Positives Fraction (TPF) is defined as percentage of Type $n > 1$ individuals receiving obituaries that contain at least one keyword:

$$\mathrm{TPF_{t,n}} = P(\mathrm{R_i} = R^+ | \mathrm{T}_i^I = n) \tag{15}$$

For a given value of n, we compare lists of different length t , and calculate the optimal length t_n^* according to the criteria of FPF, TPF and other indexes widely used in the literature: the Youden J Index, the Euclidean Distance, Accuracy index, and F-1 Score (all described below). Based on this procedure, we first compare the performance of unigrams and bigrams, and then select the optimal length of the keyword list across increasing values of n.

D.1 The ROC curve

For the simplest case, where $n = 2$, we illustrate the results using a Receiver Operating Characteristic (ROC) curve (Figure 5). This procedure is applied to both uni-grams,

green curve and bi-grams, blue curve. Each point on a curve represents the combination of (FPF, TPF) pairs that is associated to a specific number of keywords (t), increasing as we move to the right. The shape of the curve has an intuitive explanation. As we include more keywords of lower discriminating power, our the detection condition for Type $n = 2$ becomes less stringent, resulting in more positive results R^+ both for Type 1 (FPF) and Type $n = 2$ (TPF) . When we include a sufficiently large number of keywords all obituaries include at least one keyword, so every individual is tested positive and classified as Type $n = 2$.The ideal point in the figure is the upper left corner, where FPF=0 and TPF=1. A random classification obtaining by flipping a coin would produce an equal number of False and True positive as represented by the diagonal.

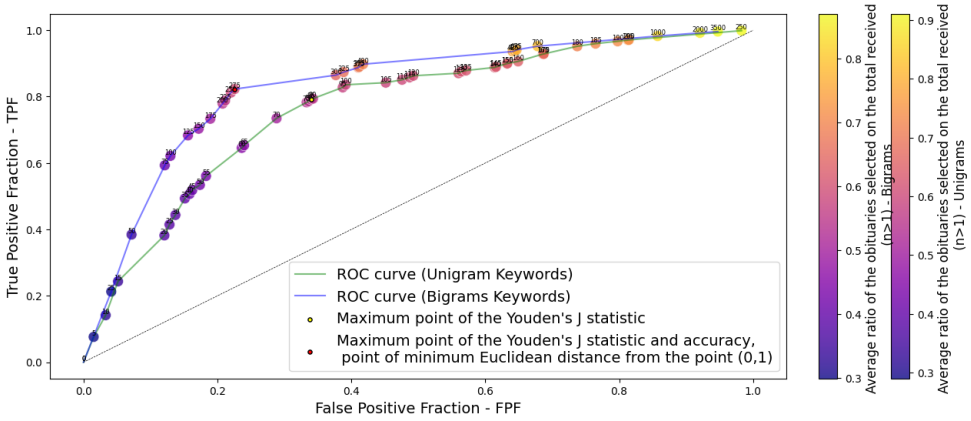

Figure 5: The Receiver Operating & Characteristic (ROC) Curve

The Figure 5 shows that all non empty lists of uni-grams and bi-grams do better (i.e. produce a larger TPF for any given FPF, or a lower FPF for given TPF) than a random classifier. Moreover, for bi-grams the closest to the ideal point is reached a list containing 275 items, while for uni-grams there is a little more ambiguity. Bi-grams always perform better, as their curve always lies above that of uni-grams.

D.2 Threshold evaluation

We iterate the procedure for increasing values of n, comparing obituaries of Type 1 and Type n individuals, and finally select the optimal list as the one that achieves the highest average score. This means that we ideally require our list to contain keywords that only to appear in obituaries for individuals who receive an increasing number of obituaries (TPF) , while appearing as little as possible among those who receive only one.

In other words, we compare keywords appearing in "single" obituaries with those appearing for individuals receiving up to $n = 15$ obituaries (thereafter results stabilize), and for each comparison we evaluate lists containing the top t NPV keywords and select the one that score best according to different criteria. We iterate this procedure separately

for bigrams and unigrams. The results are shown in the following Tables 8 and 9.

Table 8: Optimal number of keywords by n, bigrams

n	Number of optimal keywords	Criterion	FPF	TPF	Youden's J index	Euclidean Distance	Accuracy	F1-Score
1	275	Max Accuracy	0.226	0.823	0.597	0.287	0.791	0.735
		Max Youden J	0.226	0.823	0.597	0.287	0.791	0.735
		Min Euclidean Dist	0.226	0.823	0.597	0.287	0.791	0.735
		Max F1-Score	0.226	0.823	0.597	0.287	0.791	0.735
2	100	Max Accuracy	0.130	0.751	0.621	0.281	0.840	0.702
	275	Max Youden J	0.226	0.929	0.703	0.237	0.813	0.713
	225	Min Euclidean Dist	0.211	0.904	0.692	0.232	0.817	0.712
	250	Max F1-Score	0.220	0.921	0.701	0.234	0.815	0.713
3	75	Max Accuracy	0.121	0.800	0.678	0.234	0.864	0.689
	250	Max Youden J	0.220	0.965	0.745	0.223	0.815	0.663
	125	Min Euclidean Dist	0.156	0.884	0.728	0.194	0.852	0.692
	100	Max F1-Score	0.130	0.829	0.699	0.215	0.862	0.694
4	50	Max Accuracy	0.072	0.588	0.517	0.418	0.878	0.588
	125	Max Youden J	0.156	0.921	0.765	0.175	0.855	0.653
		Min Euclidean Dist	0.156	0.921	0.765	0.175	0.855	0.653
	100	Max F1-Score	0.130	0.873	0.743	0.182	0.871	0.666
5	50	Max Accuracy	0.072	0.623	0.551	0.384	0.892	0.578
	125	Max Youden J	0.156	0.949	0.793	0.164	0.856	0.611
	100	Min Euclidean Dist	0.130	0.906	0.776	0.160	0.874	0.631
	75	Max F1-Score	0.121	0.884	0.762	0.168	0.879	0.634
6	25	Max Accuracy	0.041	0.392	0.351	0.609	0.903	0.443
	125	Max Youden J	0.156	0.965	0.809	0.160	0.856	0.568
	100	Min Euclidean Dist	0.130	0.928	0.798	0.148	0.876	0.595
	75	Max F1-Score	0.121	0.909	0.787	0.152	0.882	0.601
7	25	Max Accuracy	0.041	0.418	0.376	0.584	0.915	0.442
	125	Max Youden J	0.156	0.972	0.816	0.158	0.854	0.519
	100	Min Euclidean Dist	0.130	0.941	0.811	0.143	0.876	0.551
	75	Max F1-Score	0.121	0.923	0.801	0.144	0.882	0.559
8	25	Max Accuracy	0.041	0.443	0.402	0.559	0.924	0.439
	100	Max Youden J	0.130	0.957	0.827	0.137	0.876	0.510
	75	Min Euclidean Dist	0.121	0.942	0.820	0.135	0.883	0.520
	50	Max F1-Score	0.072	0.710	0.638	0.299	0.914	0.525
9	25	Max Accuracy	0.041	0.466	0.424	0.536	0.930	0.437
	100	Max Youden J	0.130	0.968	0.838	0.134	0.876	0.477
	75	Min Euclidean Dist	0.121	0.954	0.833	0.130	0.883	0.488
	50	Max F1-Score	0.072	0.728	0.656	0.282	0.917	0.505
10	25	Max Accuracy	0.041	0.479	0.437	0.523	0.934	0.424
	100	Max Youden J	0.130	0.974	0.844	0.132	0.875	0.441
	75	Min Euclidean Dist	0.121	0.963	0.842	0.127	0.883	0.454
	50	Max F1-Score	0.072	0.742	0.671	0.268	0.919	0.480
11	25	Max Accuracy	0.041	0.501	0.460	0.500	0.938	0.418
	75	Max Youden J	0.121	0.974	0.853	0.124	0.883	0.424
		Min Euclidean Dist	0.121	0.974	0.853	0.124	0.883	0.424
	50	Max F1-Score	0.072	0.763	0.691	0.248	0.921	0.460
12	25	Max Accuracy	0.041	0.520	0.479	0.482	0.942	0.408
	75	Max Youden J	0.121	0.979	0.858	0.123	0.883	0.393
		Min Euclidean Dist	0.121	0.979	0.858	0.123	0.883	0.393
	50	Max F1-Score	0.072	0.778	0.706	0.234	0.922	0.437
13	25	Max Accuracy	0.041	0.534	0.493	0.468	0.944	0.395
	75	Max Youden J	0.121	0.981	0.859	0.123	0.882	0.363
		Min Euclidean Dist	0.121	0.981	0.859	0.123	0.882	0.363
14	25	Max Accuracy	0.041	0.556	0.514	0.446	0.946	0.384
	75	Max Youden J	0.121	0.986	0.865	0.122	0.882	0.334
		Min Euclidean Dist	0.121	0.986	0.865	0.122	0.882	0.334
	50	Max F1-Score	0.072	0.800	0.729	0.212	0.924	0.389
15	25	Max Accuracy	0.041	0.575	0.533	0.427	0.948	0.375
	75	Max Youden J	0.121	0.989	0.867	0.122	0.882	0.311
		Min Euclidean Dist	0.121	0.989	0.867	0.122	0.882	0.311
	25	Max F1-Score	0.041	0.575	0.533	0.427	0.948	0.375

Table 9: Optimal number of keywords by n, unigrams

n	Number of optimal keywords	Criterion	FPF	TPF	Youden's J index	Euclidean Distance	Accuracy	F1-Score
1	55	Max Accuracy	0.183	0.561	0.378	0.476	0.727	0.591
	85	Max Youden J	0.340	0.792	0.452	0.399	0.707	0.655
	70	Min Euclidean Dist	0.288	0.735	0.446	0.392	0.720	0.649
	85	Max F1-Score	0.340	0.792	0.452	0.399	0.707	0.655
2	15	Max Accuracy	0.052	0.305	0.253	0.697	0.788	0.418
	70	Max Youden J	0.288	0.835	0.547	0.332	0.743	0.618
		Min Euclidean Dist	0.288	0.835	0.547	0.332	0.743	0.618
		Max F1-Score	0.288	0.835	0.547	0.332	0.743	0.618
3	15	Max Accuracy	0.052	0.353	0.302	0.649	0.836	0.449
	70	Max Youden J	0.288	0.890	0.602	0.308	0.745	0.569
	65	Min Euclidean Dist	0.240	0.825	0.585	0.297	0.772	0.578
	55	Max F1-Score	0.183	0.733	0.550	0.324	0.801	0.582
4	15	Max Accuracy	0.052	0.389	0.337	0.614	0.866	0.461
	70	Max Youden J	0.288	0.911	0.623	0.302	0.741	0.510
	65	Min Euclidean Dist	0.240	0.858	0.618	0.279	0.775	0.530
	55	Max F1-Score	0.183	0.773	0.590	0.291	0.811	0.547
5	5	Max Accuracy	0.015	0.167	0.151	0.834	0.888	0.260
	65	Max Youden J	0.240	0.890	0.650	0.264	0.776	0.485
		Min Euclidean Dist	0.240	0.890	0.650	0.264	0.776	0.485
	35	Max F1-Score	0.151	0.741	0.590	0.300	0.836	0.517
6	5	Max Accuracy	0.015	0.186	0.170	0.815	0.906	0.280
	65	Max Youden J	0.240	0.908	0.669	0.257	0.775	0.442
	55	Min Euclidean Dist	0.183	0.837	0.654	0.245	0.819	0.476
	35	Max F1-Score	0.151	0.769	0.618	0.276	0.841	0.487
7	5	Max Accuracy	0.015	0.202	0.187	0.798	0.921	0.294
	65	Max Youden J	0.240	0.923	0.683	0.252	0.773	0.397
	55	Min Euclidean Dist	0.183	0.856	0.673	0.233	0.820	0.435
	15	Max F1-Score	0.052	0.491	0.440	0.511	0.911	0.473
8	5	Max Accuracy	0.015	0.222	0.207	0.778	0.933	0.311
	65	Max Youden J	0.240	0.932	0.692	0.249	0.772	0.355
	55	Min Euclidean Dist	0.183	0.870	0.687	0.225	0.821	0.395
	15	Max F1-Score	0.052	0.511	0.459	0.492	0.919	0.459
9	5	Max Accuracy	0.015	0.237	0.222	0.763	0.941	0.320
	65	Max Youden J	0.240	0.941	0.701	0.247	0.771	0.324
	55	Min Euclidean Dist	0.183	0.883	0.700	0.217	0.821	0.365
	15	Max F1-Score	0.052	0.526	0.474	0.477	0.924	0.446
10	5	Max Accuracy	0.015	0.256	0.241	0.744	0.948	0.332
	55	Max Youden J	0.183	0.894	0.711	0.211	0.821	0.335
		Min Euclidean Dist	0.183	0.894	0.711	0.211	0.821	0.335
	15	Max F1-Score	0.052	0.546	0.494	0.457	0.928	0.434
11	5	Max Accuracy	0.015	0.275	0.259	0.726	0.953	0.343
	55	Max Youden J	0.183	0.904	0.721	0.206	0.821	0.309
	45	Min Euclidean Dist	0.162	0.874	0.712	0.205	0.839	0.325
	15	Max F1-Score	0.052	0.568	0.516	0.435	0.932	0.423
12	5	Max Accuracy	0.015	0.289	0.274	0.711	0.958	0.347
	55	Max Youden J	0.183	0.912	0.729	0.203	0.821	0.283
	45	Min Euclidean Dist	0.162	0.884	0.722	0.199	0.839	0.299
	15	Max F1-Score	0.052	0.585	0.533	0.418	0.934	0.408
13	5	Max Accuracy	0.015	0.303	0.287	0.698	0.961	0.350
	55	Max Youden J	0.183	0.915	0.732	0.202	0.820	0.259
	45	Min Euclidean Dist	0.162	0.889	0.727	0.196	0.839	0.275
	15	Max F1-Score	0.052	0.594	0.542	0.409	0.936	0.390
14	5	Max Accuracy	0.015	0.315	0.300	0.685	0.965	0.349
	55	Max Youden J	0.183	0.922	0.739	0.199	0.820	0.236
	35	Min Euclidean Dist	0.151	0.882	0.730	0.192	0.850	0.261
	15	Max F1-Score	0.052	0.615	0.563	0.389	0.938	0.375
15	5	Max Accuracy	0.015	0.321	0.306	0.679	0.967	0.344
	55	Max Youden J	0.183	0.926	0.743	0.197	0.820	0.218
	35	Min Euclidean Dist	0.151	0.891	0.740	0.186	0.850	0.243
	15	Max F1-Score	0.052	0.638	0.586	0.366	0.940	0.365

Definitions of the evaluation metrics used in this process:

- The *Youden's J Index* is defined as the difference between the True Positive Fraction and the False Positive Fraction. For a given length t and for a given n, the Youden's J Index is equal to:

$$\mathrm{YJI_{t,n}} = \mathrm{TPF_{t,n}} - \mathrm{FPF_{t,1}} \tag{16}$$

  This index is particularly well suited for our application, as it handles unbalanced data by being independent of the number of observations classified as 'Type 1' and 'Type n'.

- The *Euclidean Distance* refers to the distance between each pair $(\mathrm{FPF}_{t,1}, \mathrm{TPF}_{t,n})$ and the point of optimal performance $(0, 1)$ in the ROC space. For a given length t and for a given n, the Euclidean Distance is equal to:

$$\mathrm{ED_{t,n}} = \sqrt{(\mathrm{FPF}_{t,1})^2 + (1 - \mathrm{TPF}_{t,n})^2} \tag{17}$$

- The *Accuracy Index* is defined as the proportion of correct predictions out of the total number of predictions. In this application, accuracy corresponds to the proportion of individuals correctly classified as both 'Type 1' and 'Type n' out of the total number of individuals. The *Accuracy* for given t and n is:

$$\mathrm{Accuracy}_{t,n} = P(\mathrm{T}_i^I = 1) \times \left(1 - \mathrm{FPF}_{t,1}\right) + P(\mathrm{T}_i^I = n) \times \mathrm{TPF}_{t,n} \tag{18}$$

  Since this metric is influenced by the relative size of the two categories, it can lead to misleading conclusions in presence of unbalanced data.

- The *F1 Score* is defined as the harmonic mean of Precision and Recall, where *Recall* is equal to the True Positive Fraction, whereas *Precision* is defined as the ratio between the number of 'Type n' individuals whose obituaries contain at least one of our ranked keywords and the total number of individuals (both 'Type 1' and 'Type n') whose obituaries contain at least one of our ranked keywords.
  The *F1 Score* for given t and n is:

$$\mathrm{F1Score}_{t,n} = 2 \times \frac{\mathrm{Precision}_{t,n} \times \mathrm{TPF}_{t,n}}{\mathrm{Precision}_{t,n} + \mathrm{TPF}_{t,n}} \tag{19}$$

  where $\mathrm{Precision}_{t,n} = \frac{P(\mathrm{R_i}=R^+ \cap \mathrm{T}_i^I = n)}{P(\mathrm{R_i}=R^+)}$
  When the negative class is much larger than the positive class, even a small false positive rate can produce more false positives than true positives in absolute numbers. This reduces precision, despite a high recall.

Appendix E: Random Forest and Support Vector Machine

Random Forests, introduced by Breiman (2001), are a combination of multiple independent classification trees. A supervised classification algorithm splits the dataset into homogeneous predefined classes based on iterative tests such as the presence or absence of a particular feature. Beginning at the root node, where no division has yet occurred, each step selects a splitting rule that best separates the labeled data into the two classes, according to a criterion, such as information gain, chi-square, or the Gini index.

The trees continue subdividing the data until they reach terminal nodes, where stopping criteria are met, such as achieving node purity or reaching a predefined maximum tree depth (Quinlan, 1986, Ali et al., 2012, Mienye & Jere, 2024). A key advantage of this algorithm is the ability to provide a measure of feature "importance", which enhances their interpretability and makes them particularly valuable for economic applications (Loh, 2014).

Examples of Random Forest application to Natural Language Processing (NLP) techniques include Elagamy et al. (2018), who used financial news articles to isolate unigrams and bigrams in order to predict Dubai's stock market changes. Similarly, Mao et al. (2023) used the feature importance scores of Random Forest to compare their text-based technological innovation index with traditional financial indicators in credit risk prediction.

Support Vector Machines, introduced by Boser et al. (1992), are supervised algorithms that achieve classification based on a graphical representation of observations in a feature space. During training, the algorithm selects the hyperplane that best separates the labeled observations into the different classes. This is achieved by maximizing the margin, which is defined as the distance between the hyperplane and the support vectors (I.e., the observations closest to the hyperplane).

In the following two sections of this appendix, we provide details on the configuration and training of the models, as well as on their classification performance and validation.

E.1 Configuration and Training

We trained the Random Forest Classifier and the Support Vector Machine provided by the Scikit-Learn library. The dataset used for training and evaluation comprised 5,062 manually labelled obituaries, which were randomly divided into a training set (90%) and a test set (10%). These texts were vectorised in a binary term-frequency matrix, where the features considered were restricted to ourlist of keywords comprising the top 275 bigrams and the top 85 stemmed unigrams, selected according to the NPV values. This implies that the classifiers receive as input a matrix having a row for each obituary in our corpus and a column for each of the 360 selected keywords. An entry in the matrix is 1 if the corresponding keyword is present in the obituary, and 0 otherwise. By considering solely the presence or absence of features, we ensure that all instances of keywords are treated equally when we classify. We also experimented with training classifiers on feature counts and term frequency–inverse document frequency representations and found classification results to be very similar across these vectorisation methods.

To improve generality and avoid overfitting, we trained both the Random Forest and the Support Vector Machine models using a 15-fold cross-validation model. Therefore, the training process was repeated fifteen times, with fourteen folds used for training and one for validation in each iteration. This ensured that each fold was used for both training and validation. This method enabled us to identify the optimal parameters based on cross-validation results.

Random Forest and Support Vector Machine hyper-parameters were determined using the Grid Search method provided by the Scikit-Learn library.

For the Random Forest, the number of decision trees was chosen from 100, 300, 500, and 1000, allowing us to balance the trade-off between computational efficiency and classification performance across models of varying complexity. The criterion for splitting nodes was based on Gini index, prioritising the creation of homogeneous subsets. The Gini index provides a measure of purity for each node of the tree, reaching zero when maximum purity is achieved — that is, when all observations in the node belong to the same class. The maximum depth of the trees was tested with an unlimited depth as well as constraints of 10 and 20. The minimum number of samples required to split an internal node was choosen from 2, 5 and 10, while the minimum samples per leaf was selected from 1,2 and 4, allowing the trees to expand without early stopping. Finally, to make the trees learn different patterns from the data, we restricted the number of features that could be considered at each node for the splitting criterion. The model selected this number from three options: the square root, the base-2 logarithm, or half of the total number of features. The 15-fold cross-validation process selected as the optimal model the one consisting of 100 trees, with unlimited maximum depth. Each internal node had to have at least 10 samples to be split, and each leaf node had to contain at least one sample. The optimal number of features to consider in the splitting was given by the base-2 logarithm of the total number of features, which is around 8 or 9 features ($\log_2(275 + 85)$).

For the Support Vector Machine, we tested different values for three key hyperparameters: the regularization parameter, the kernel function, and the kernel coefficient. These are described in the Scikit-Learn library documentation. The regularization (C) parameter is used to handle the trade-off between minimizing classification errors during the training phase and maximizing the margin, which ensures better generalizability in out-of-sample classification. A larger value of C allows for a smaller margin if it results in better classification accuracy on the training data. We tested this parameter with values equal to 0.1, 1, 10, and 100. The kernel types determine the feature space in which both the observed data and the separating hyperplane are represented. Kernel function allow us to project observation data points into a higher-dimensional feature space. This enables the SVM to find a linear separation in the transformed space also when the data is not linearly separable in the two dimensional space. We tested three common kernels: linear, radial basis function (RBF), and polynomial. The kernel coefficient (gamma) determines the influence of a single training example in the feature space. When gamma is high, this influence is very localized, leading to complex decision boundaries and the risk of overfitting. Conversely, when gamma is low, each point also influences very distant

points, leading to underfitting and lower classification performance. We tested for the gamma parameter with the values "scale", "auto", 0.1 and 1. The 15-fold cross-validation process selected the model based on the radial basis function (RBF) kernel as the optimal one, with the regularization parameter set to 0.1 and the gamma parameter equal to 1. The radial basis function kernel maps the data into a higher-dimensional space by computing the Euclidean distance between pairs of vectors representing observation points, weighted by the gamma parameter. This allows the model to find linear relationships and select the hyperplane that maximizes the margin.

The figure below (Figure 6) shows the most important features used by the Random Forest for classifying obituaries. These were obtained using the "*feature_ importances_*" attribute from Scikit-Learn's Random Forest Classifier.

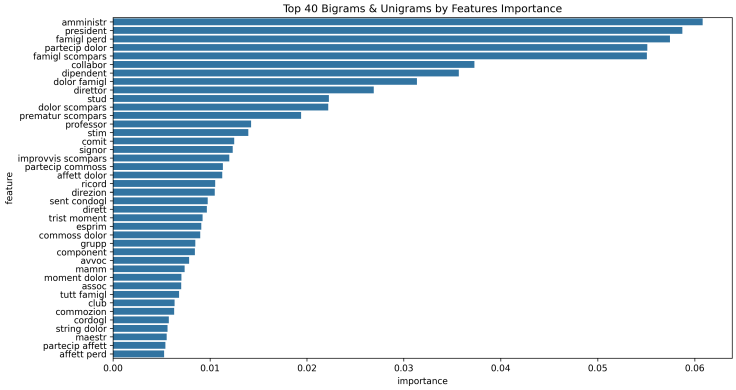

Figure 6: Top 40 Unigrams and Bigrams by Random Forest Features Importance Score

Figure 6 shows that the two most important stems for classification are associated with the professional sphere (e.g., "*amministratore*" which means administrator, "*presidente*" which means president). While general expressions of condolence are also present, a substantial share of the top 40 stems used by the trained Random Forest classifier are profession-related.

E.2 Classification and Perfomance Validation

Tables 10 and 11 show the out-of-sample prediction performance of the Random Forest and Support Vector Machine models. This performance is evaluated by classifying the obituaries in the test set (10% of the manually labeled obituaries) using the two trained classifiers and comparing the predicted labels with the manually assigned ones. The test set consisted of 507 obituaries: 306 labeled as "family" and 201 labeled as "non-family".

Table 10: Random Forest classification performance

Class	Precision	Recall	F1-score	Support
family	0.90	0.94	0.92	306
non family	0.90	0.84	0.87	201
accuracy		0.90		507
macro avg	0.90	0.89	0.89	507
weighted avg	0.90	0.90	0.90	507

Table 11: Support Vector Machine classification performance

Class	Precision	Recall	F1-score	Support
family	0.92	0.87	0.89	306
non family	0.81	0.88	0.84	201
accuracy		0.87		507
macro avg	0.86	0.87	0.87	507
weighted avg	0.88	0.87	0.87	507

The Random Forest shows better overall prediction performance compared to the Support Vector Machine, achieving an accuracy equal to 90% (compared to 87%). However, both models have their own advantages when predicting non-family obituaries. The Random Forest achieves higher precision (around 90% versus 81%), while the Support Vector Machine achieves higher recall (about 87% versus 84%). This means that while the Support Vector Machine correctly labels as non-family a higher fraction of the manually labeled non-family obituaries, the Random Forest has the advantage of a higher fraction of correctly predicted non-family obituaries out of the total predicted as non-family. In other words, the SVM provides a higher true positive fraction, while the RF provides a lower false positive fraction. The description of the metrics used in this validation phase is provided at the end of this appendix.

The implications of these differences are also evident when moving from the classification of obituaries to the classification of individuals. We define an individual as "popular" if at least half of their obituaries are classified as "non-family".

Table 12 compares the classification outcomes of individuals as "popular" or "non popular" based on the obituary classifications produced by the three models: Naive, Random Forest, and Support Vector Machine. Each column shows the pair of Classifiers being compared (e.g., "*Random Forest A vs Support Vector Machine*"), each row indicates the Type label according to the two classifiers (e.g., "*(Non popular type , Popular type)*"). Each cell reports the number of individuals classified as first Type by the first classifier and as second Type by the second classifier. For example, the cell corresponding to column "RF vs Naive" and row "Non Pop, Pop" indicates that 4,769 individuals are classified as "non popular" by the Random Forest while being classified as "Popular"

by the Naive model. In this case the Naive model tends to classify as Popular 4769 more deceased than Random Forest. The last row summarizes the total frequency when the individuals are classified similarly for the different pairs of classifiers. Our favorite classifiers, RF and SVM agree on 95 percent of individuals, and the lowest agreement percentage, 90.3, is found between RF and Naive. Disagreement is due to the fact that, as expected, the Naive model is the most prone to classify individuals as popular, followed by SVM. Indeed, the Naive criterion is quite lax: an obituary is classified as non-family if only it contains one of the selected stems.

Table 12: Comparison of classification outcomes across algorithms

Classification pairs	**RF vs Naive**	**RF vs SVM**	**SVM vs Naive**
(Non pop, Non pop)	39343	41902	39343
(Non pop, Pop)	4769	2210	2564
(Pop, Non pop)	0	5	0
(Pop, Pop)	4928	4923	7133
Total Agreement (%)	**90.3**	**95.5**	**94.8**

Notes: "*Pop*" denotes individuals for whom more than **50%** of their obituaries are classified as non-family by the respective algorithm; "*Non pop*" otherwise.
"*Total Agreement*" is the percentage of individuals for which each algorithm pair produces identical classifications.

In fact, note that in the third row "Pop, Non Pop" there are no individuals that RF or SVM classifies as "popular" while being classified as "non popular by the Naive algorithm. In turn, Naive classifies as "popular" 4769 more individuals than RF, and 2210 more than SVM. The Random Forest is the most restrictive model, labeling only 4,928 individuals as popular out of the 9,697 identified by the Naive model, while the Support Vector Machine confirms 7,133 individuals as popular.

The second column shows that only 2,215 individuals out of 49,040 are classified differently by the Random Forest and the Support Vector Machine, mostly because the Random Forest is more conservative in classifying obituaries as non-family and, consequently, individuals as popular.

Metrics description:

1. *Precision*: The fraction of correctly classified obituaries in a given category out of the total number of obituaries predicted to belong to that category.
2. *Recall*: The fraction of correctly classified obituaries in a given category out of the total number of actual obituaries in that category (equivalent to the True Positive Fraction).
3. *F1-Score*: The harmonic mean of Precision and Recall.
4. *Accuracy*: The fraction of correct predictions out of the total number of predictions.
5. *Macro Average*: The average of each metric's values for obituaries labeled as family and non-family.
6. *Weighted Average*: The average of each metric's values for obituaries labeled as family and non-family, weighted by their respective counts.

References

Alfano, Mark, Higgins, Andrew, & Levernier, Jacob (2018). "Identifying Virtues and Values Through Obituary Data-Mining". In: *The Journal of Value Inquiry* 52.1, pp. 59–79. ISSN: 1573-0492. DOI: 10.1007/s10790-017-9602-0. URL: https://doi.org/10.1007/s10790-017-9602-0.

Ali, Jehad, Khan, Rehanullah, Ahmad, Nasir, & Maqsood, Imran (Sept. 2012). "Random Forests and Decision Trees". In: *International Journal of Computer Science Issues(IJCSI)* 9.

Ash, Elliott & Hansen, Stephen (2023). "Text Algorithms in Economics". In: *Annual Review of Economics* 15.1, pp. 659–688. DOI: 10.1146/annurev-economics-082222-074352.

Boser, Bernhard E., Guyon, Isabelle M., & Vapnik, Vladimir N. (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 144–152. ISBN: 089791497X. DOI: 10.1145/130385.130401. URL: https://doi.org/10.1145/130385.130401.

Bourdieu, Pierre (1988). *Homo Academicus*. Cambridge: Polity Press, pp. 216–225.

— (1996). *The State Nobility*. Cambridge: Polity Press, pp. 40–53.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Corneo, Giacomo & Jeanne, Olivier (1997). "Conspicuous consumption, snobbism and conformism". In: *Journal of Public Economics* 66.1, pp. 55–71. URL: https://EconPapers.repec.org/RePEc:eee:pubeco:v:66:y:1997:i:1:p:55-71.

Elagamy, Mazen Nabil, Stanier, Clare, & Sharp, Bernadette (2018). "Stock market random forest-text mining system mining critical indicators of stock market movements". In: *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pp. 1–8. URL: https://api.semanticscholar.org/CorpusID:46974642.

Fowler, Bridget & Bielsa, Esperança (2007). "The lives we choose to remember: a quantitative analysis of newspaper obituaries". In: *The Sociological Review* 55.2, pp. 203–226. DOI: https://doi.org/10.1111/j.1467-954X.2007.00702.x.

Li, Kai, Mai, Feng, Shen, Rui, & Yan, Xinyan (July 2020). "Measuring Corporate Culture Using Machine Learning". In: *The Review of Financial Studies* 34.7, pp. 3265–3315. DOI: 10.1093/rfs/hhaa079.

Loh, Wei-Yin (2014). "Fifty Years of Classification and Regression Trees". In: *International Statistical Review* 82.3, pp. 329–348. DOI: https://doi.org/10.1111/insr.12016. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12016. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12016.

Lorenzetto, Stefano (Feb. 8, 2015). *Ha raccolto 100mila necrologi: «Incredibile, nessuno muore»*. URL: https://www.ilgiornale.it/news/politica/ha-raccolto-100mila-necrologi-incredibile-nessuno-muore-1090976.html (visited on 06/29/2025).

Loughran, Tim & Mcdonald, Bill (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks". In: *The Journal of Finance* 66.1, pp. 35–65. DOI: https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Manasse, Paolo (2003). *Il Mistero del Necrologio Mancante*. La Voce. URL: https://www.lavoce.info/archives/21910/il-mistero-del-necrologio-mancante/ (visited on 06/29/2025).

Mao, Yang, Shifeng, Liu, & Daqing, Gong (2023). "A Text Mining and Ensemble Learning Based Approach for Credit Risk Prediction". In: *Tehnički vjesnik 30, br. 1*, pp. 138–147. URL: https://doi.org/10.17559/TV-20220623113041.

Mienye, Ibomoiye Domor & Jere, Nobert (2024). "A Survey of Decision Trees: Concepts, Algorithms, and Applications". In: *IEEE Access* 12, pp. 86716–86727. DOI: 10.1109/ACCESS.2024.3416838.

Papi, Giacomo (July 2020). *Nei necrologi si racconta il passaggio della storia, il mutare dei gusti e dei valori*. URL: https://www.ilfoglio.it/cultura/2020/07/26/news/nei-necrologi-si-racconta-il-passaggio-della-storia-il-mutare-dei-gusti-e-dei-valori-322622/ (visited on 06/29/2025).

Quinlan, J. Ross (1986). "Induction of Decision Trees". In: *Machine Learning* 1, pp. 81–106. URL: https://api.semanticscholar.org/CorpusID:189902138.

Shapiro, Adam Hale, Sudhof, Moritz, & Wilson, Daniel J. (2022). "Measuring news sentiment". In: *Journal of Econometrics* 228.2, pp. 221–243. DOI: https://doi.org/10.1016/j.jeconom.2020.07.053.

Veblen, Thorstein (1918). *The Theory of the Leisure Class: An Economic Study of Institutions*. New York: B. W. Huebsch.

Alma Mater Studiorum - Università di Bologna

DEPARTMENT OF ECONOMICS

Strada Maggiore 45
40125 Bologna - Italy
Tel. +39 051 2092604
Fax +39 051 2092664
http://www.dse.unibo.it