# Venture Capital Firm Linkage and Document Similarity

**Authors**
Vikram Sharma
Maya Sijaric
Stefan Mrakovcic

## Abstract

In an effort to establish the most optimally performing similarity algorithms for the identification of likely early-stage investment targets for Venture Capital (VC) funds, we surveyed (1) the Jaccard Similarity (Baseline), (2) Cosine Similarity, (3) Traditional TF-IDF, (4) Word-Mover's Distance (ML-based), and (5) Sent2Vec (ML-based) algorithms. On the basis of Crunchbase-derived sets of textual descriptions of funded startups from VC portfolios, we found that the high degree of similarity within our data, and unquantifiable events that make up real-life investment decisions, complicates efforts to build a trained model capable of matching portfolio companies to their respective VC partners. Nevertheless, by tightening our threshold value and by pursuing less ML-driven methods for establishing similarity, we observed an increase in the prediction accuracies, making this methodology suitable for attempts to meaningfully match startups to appropriate VC portfolios. The deeper question of what makes a good "match", however, must be further researched, if a more complete answer to our research question is to be reached.

## Introduction

Entrepreneurs commonly face two major roadblocks in their pursuit of financing: information overload and information restrictions. In a similar vein, it is clear that VCs' pursuits of investment candidates, whose entrepreneurial endeavors and needs ought to align with their expertise and mission, also face similar challenges, which might be alleviated by a similarity-based computerized approach to investment decisions and financing.

1. **Information Overload:** The sheer number of VC firms operating in the United States proves to be the first challenge. Despite this plethora of VCs and the abundance of financial resources available in the United States, investment decisions of this nature are far more nuanced and one is highly unlikely to stumble upon the right VC in a reasonable timescale - due diligence can only get one so far (Pit). These decisions rely and extend beyond the mere existence of fundable enterprises and skilled teams, and involve a myriad of other requisites. In consideration of potential investment candidates, questions that arise in VC settings include "is this the right industry for us?", "is this team in the appropriate funding stage?" and "are they located in a strategically optimal market, and/or with respect to us?" - questions which might be more efficiently answered by means of a similarity-based algorithm.
2. **Information Restriction:** The problem posed by paywalls on financial databases, and lack of connections/networks, might hinder an otherwise excellent entrepreneur from successfully raising the necessary capital needed for their operation. The VC domain is characterized by a level of ingrained exclusivity, which has kept the inexperienced founder out of consideration and has thereby killed many viable ideas in their infancy. The "insider knowledge" needed to not only reach, but also know whom to reach out to, is fundamental in this regard.

With the overarching goal of alleviating these two concerns, we collected data on VC portfolios from Crunchbase. More specifically, and perhaps more elegantly put, we gathered descriptions of companies that had been funded by VCs as well as the fund that had invested in them. After splitting the data into a development, training, and testing set, we surveyed 5 algorithms attempting to establish how similar randomly selected, previously-funded companies would be compared to all the available VCs' portfolio companies, in order to find the portfolio that the company is most likely to belong to. This methodology serves to identify "true matches" or "best guesses," if you will, and lays the foundation for a probabilistic recommendation engine that has the potential to aid the complex decisions described above. At the outset, we hypothesized that the more involved ML-based algorithms would develop a better understanding of the texts and that it would hence outperform the more traditional NLP approaches, which would make them more meaningful indicators of similarity.

## Related Literature - Financial Domain

The primary research paper we studied in preparation for our endeavor, outlined Latent Variable Models (Gong et al., 2018), which are domain-specific document probability-based and count-based models for establishing similarity. It also involves a consideration of Word Movers' Distance, Word Embeddings (such as systems powered by Doc2Vec and Word2Vec), as well as baseline similarity methods, such as Cosine similarity and Jaccard similarity. This helped us establish some common baseline systems, which we implemented in our inquiry into the similarity between child companies. Term/data mismatch was also brought to our attention by this paper and it was presented as an issue when comparing documents to their respective summaries, which served to caution us about a potential problematic encounter along the same lines. In our case, however, this concern would, of course, be more relevant in the event that we found the child companies to be so different as to make their comparisons unproductive, possibly due to their representation of different industries, for example. To our surprise, we found the descriptions of the child companies to, instead, be very similar to one another, despite the fact that they belong to many vastly different industries, such as cloud computing, infrastructure, VR, and eCommerce.

A process of summarizing financial documents by the University of Madrid had the goal of making them easier and faster to comprehend, as to avoid reading the entire document, one would read a summary generated from the original (Baldeon Suarez et al., 2020). The application of TF-IDF for the evaluation of the relevance of words, as well as a number of techniques for pre-processing input data, were introduced in the authors' efforts to enhance summarization methodologies for financial documents and to produce their "gold-standard summary". This encouraged our team to use similar techniques, with respect to our own dataset and research question, given that this would help us place a greater emphasis on the more important words and key points in the child company description. More specifically, we sought to prepare the data for the subsequent surveying of similarity algorithms, by removing stop words, utilizing lemmatization, and lowercasing the dataset.

## Related Literature - General Similarity

The researchers at Northwestern and Intel set out to analyze different versions of a document over the course of its editing processes and then use common metrics to evaluate their similarity models (Zhu et al., 2017). While our research may focus on different topics, the standard of evaluating the different models was achieved by defining a custom value of "goodness" for any distance measure between documents, and then using the ground truths of a "good match" to compare it to the match produced by the system when calculating precision, recall, and F-measure. This technique of customized "good matches", combined with our three commonly associated metrics provided us with a more in-depth understanding of our findings.

An additional article laid the groundwork for the initial stages of our project as our research formulated around a similar method. The researchers from Europe created an improved method of citation ranking for ranking individual research papers against a set of scholarly submissions (Leydesdorff et al., 2011). This encouraged us to model our methodology by comparing the descriptions of individual child companies against sets of descriptions in individual portfolios. Moreover, it suggests possible next steps for our own research, primarily, the possibility of an ensemble learning model that combines multiple similarity metrics for prediction optimization purposes.

## Data

The data was derived from Crunchbase.com, a financial services database that contains information on VCs and various details on their investments. Access to these details was made possible by a paid subscription to the website, with the intention of studying the data for non-commercial purposes. We targeted and extracted data relevant to each selected VC firm, all of which are located in the United States. This data includes, most importantly, the textual descriptions of their portfolio companies, which are provided in the form of paragraphs that average 80-120 words in length. This extraction resulted in CSV files that laid the groundwork for our subsequent inspection and study of the relationships between VCs ("parent company") and their portfolio companies ("child company").

Some of the most reputable VC firms, widely recognized to be amongst the top 30 highest performing, were picked in order to widen our scope of included industries. Our research was driven by the operating assumption that well-funded firms, as well as legacy firms, are more likely to be better equipped with the resources and knowledge necessary to invest agnostically across more diverse sectors in the market, thereby analyzing the investments of these firms provides us with access to a broader range of industries, which are typically not available to niche firms, such as a MedTech-oriented fund that only invests in healthcare technology, for instance. In order to further broaden the scope of our data, we included all their child companies, some of which represented an investment that can be traced back to the1990s. For the parent companies, we took in information specifying a set that represents their most recent 1000 investments, or fewer, if it was the case that they had made fewer than 1000 investments. For the child companies in the portfolios, we took in the following data: the parent company, a full description of the child company, their primary industry involvements, their funding stage, and the most up-to-date current funding amount acquired.

The data collected was intended to provide us with as much unbiased information as possible, in order to meaningfully account for a multi-faceted industry. It is possible that this decision might have inadvertently introduced an unintended bias. The parent companies chosen were from the upper echelons of the business financing world, as we sought to acquire a dataset encompassing as many industries as possible. We made the assumption that the child companies of these top VCs were comparable to the child companies of the bottom tier of VCs; while this might be the case, further analysis is warranted for such a proclamation.

We wrote a python script that read in these details about the child companies from the original CSV files, such that we were left with the most relevant details and descriptions that had been processed using three methods. Firstly, we removed any occurrence of stop words from the child company descriptions, followed by the lower casing of said text, as well as the application of lemmatization techniques, in order to improve the efficiency and accuracy of our surveys, and, more importantly, create representations of the text that better capture the distinguishing features of each child company.

## Methodology

The data was split up into three sets: $80\%$ for the training set, $5\%$ for the development set, and $15\%$ for the testing set. The development set was then used to determine the threshold value for the evaluation that followed. We learned that when the threshold was increased from the top $15\%$ to instead reflect outputs based on the top $50\%$, the system predicted more overall investors as "good matches". Due to the fact more accurate predictions were found to be more meaningful in the applications of our system, a proposition justified in the "Discussion" section, it appears that the trade-off between accuracy and recall ought to be balanced in a manner that prioritizes higher accuracy scores. Accuracy scores were derived using the absolute number of False Negatives, False Positives, True Negatives, and True Positives and were checked against the answer key, i.e. the reality of which investors actually invested in that specific child company. Lower accuracy scores, coupled with higher recall scores, would have rendered our system devoid of useful real-life applications.

We surveyed five existing algorithms, in order to determine their relative performances, in terms of their success when attempting to identify our pre-established VC-startup matches, on the basis of average similarity. Note that we define average similarity to mean the average of all the similarities between individual child company descriptions and all the entries ("child company descriptions") within a VC's portfolio in our training set. These algorithms included: (1) the Jaccard Similarity (Baseline), (2) Cosine Similarity, (3) Traditional TF-IDF, (4) Word-Mover's Distance (ML-based) and (5) Sent2Vec (ML-based). Thereafter, 24000 training entries and 4500 testing entries were used to determine the precision, recall, f-score, and accuracy scores using different thresholds and traditional multi-class metric evaluation techniques defined below.

- $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- $precision = \frac{TP}{TP+FP}$
- $recall = \frac{TP}{TP+FN}$
- $f1 = \frac{2*precision*recall}{precision+recall}$

## Results

Two patterns in the results of our surveys stand out and occupy the forefront of our attention. Firstly, we noticed that nearly all the algorithms produced relatively high accuracy scores when the threshold was changed from top $50\%$ to top $15\%$. These findings were coupled with less impressive recall scores, however, which suggests that our documents were likely homogeneous to some degree.

Secondly, we noted that Sent2Vec and Word-Mover's Distance, the two ML-focused approaches surveyed in this paper, were outperformed by the TF-IDF method, as well as the Cosine similarity and the Jaccard similarity algorithms, when identifying similarities between the individual descriptions of child companies and individual portfolios, i.e. sets of VCs' child company descriptions, when the algorithms attempt to correctly determine which portfolio it is most likely to belong to, i.e. be "most similar to". Even though the ML-based approaches resulted in fewer correct matches between individual companies and portfolios, and despite the fact that they did not display a clear advantage over the other techniques, in this regard, we do not yet have a basis to make any grand claims. As this paper has alluded to in our "Data" section, it is possible that our descriptions did not paint a sufficiently complete picture of the reality faced by startup founders and VCs. Perhaps more data of this kind, or another type of data altogether, would have produced the opposite outcome than we initially anticipated.

We observed an improvement in the F1 scores for all the algorithms when the threshold was changed from the more restrictive figure of the top $15\%$ to the top $50\%$. Interestingly, the ML-focused algorithms improved the most by far - the F1 score of Word-Mover's Distance, specifically, increased from $\sim 0.032$ to $\sim 0.142$. Indeed, in this regard, the ML-focused algorithms outperform traditional metrics, when the threshold was changed from top $15\%$ to top $50\%$.

In light of the observations about F1 scores outlined above, a deeper inspection of the somewhat surprising accuracy scores that followed is warranted. While all the similarity algorithms performed better in environments dictated by a more restrictive threshold (i.e. the top $15\%$), the ML-focused similarity algorithms were indeed outperformed by the traditional approaches, regardless of the threshold values used. Notably, the better performing ML model, sent2vec, had an accuracy score of $\sim 0.7454$, when the threshold was defined as the top $15\%$, which was $\sim 0.1$ less than that of the traditional approaches' accuracy scores. However, sent2vec's accuracy score displayed a greater resiliency, in the face of a threshold change from top $15\%$ to top $50\%$, suggesting that one might observe a decent performance with ML-models, in terms of accuracy, when there is a tighter threshold coupled with data of an alternative kind.

In terms of recall, all the models surveyed were marked by better scores, when the threshold was reduced from top $15\%$ to top $50\%$. There was, however, no change in relative performances, when this change occurred. It is worth noting that the traditional approaches' recall benefited more from this change to a less restrictive threshold. Jaccard's Similarity algorithms' recall score improved by $\sim 0.2908$, from $\sim 0.17658$ to $\sim 0.46738$, for example, whereas the recall score of the better performing ML-focused algorithm, sent2vec, was marked by a more modest change of $\sim 0.1603$, from $\sim 0.04545$ to $\sim 0.2058$.

Finally, in consideration of the relative precision score changes induced by a change in the threshold, from the conservative figure of top $15\%$ to top $50\%$, it is evident that the precision of all the similarity algorithms surveyed improved. The changes in precision observed with the ML-focused algorithms, nevertheless, stand out. The improvement in the precision of the Word-Mover's Distance algorithms, in particular, was marked by an increase of $\sim 0.1$, when the threshold was loosened, whereas the more traditionally oriented techniques, instead, improved by a mere $\sim 0.01$ on average. More specifically, the precision of Jaccard's Similarity algorithm, for instance, improved from $\sim 0.0470$ to $\sim 0.0570$.

## Discussion

The results outlined above, naturally, lead us to the question of "which metrics matter the most?" In light of the trade-off between precision and accuracy specified in Figure 1, one could make the argument that similarity-based approaches to matching startups candidates to suitable VCs would be of greatest utility when the models produce a few accurate, high-quality matches, rather than more "precise" matches, as these would not be hugely informative when startups are trying to navigate to their ideal VC in a vast ocean of "wrong" VCs. This conjecture is particularly reasonable when viewed through the lens of the initially defined concerns regarding "information overload". Likewise, an emphasis on recall, as opposed to accuracy, would derail us from our goal of alleviating the problem of information overload, as compromising accuracy for the sake of recall would cause the models to produce more matches, but a number of them would be tainted by low accuracy, making them less useful, given that our primary objective was to help identify methods for lessening information overload concerns. For instance, if this research were to be translated into real-life applications in the world of entrepreneurial financing, it appears reasonable to construct models whilst keeping in mind the operating assumption that founders would be more interested in a shortlist of promising ("more accurate") recommendations.

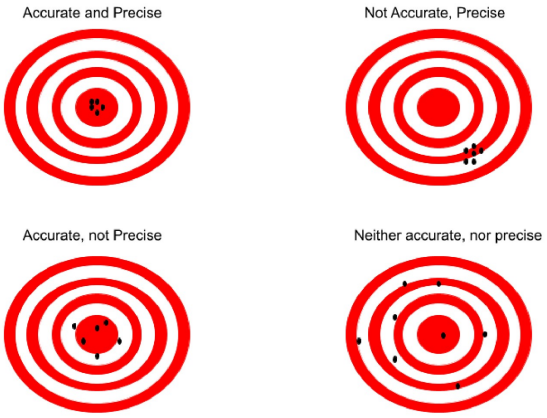


Figure 1: Relationship between accuracy and precision

The lower accuracy of matches resulting from the surveyed ML-focused models must also be addressed, given the possibility that real-life extenuating circumstances complicate our foundational definition of what a "good match between a startup and a VC" actually entails. An accurate reflection of reality, potentially facilitated by means of a tight threshold and the less ML-intense algorithms, does not necessarily indicate an ultimate ideal to begin with, given that randomness likely compromises the ability of VCs and startups to properly identify one another in the real world. The ML-based models' lower accuracy should not necessarily be understood as a poor match, but rather a reflection of the randomness inherent in investment decisions, such as short-term capital shortage. In other words, ML has no clear advantage over traditional NLP techniques, if our objective is to identify patterns in the real world. However, if the objective is to identify the most promising matches by including and picking up on potentially overlooked patterns in the portfolios and child company descriptions, then a deeper inquiry would be needed to establish such a potential utility. The inclusion of more, complementary, or simply new data altogether might potentially alter the dynamics of the accuracies produced by the ML-oriented algorithms surveyed here.

As was alluded to in the introductory portion of the "Results" section, we identified a high degree of internal similarity across the child company descriptions, which might explain the lower F1 scores observed for all these algorithms. It is possible that many founders frequently utilized buzzwords, and other specific words, in their descriptions, possibly with the unconscious, or even conscious, intention of attracting VC funding in this manner. Since this would imply that the resulting lack of emphasis on distinguishing/keywords in these descriptions would compromise the efficiency of classifications needed to produce high recall scores, it would likely be a worthwhile research endeavor to tinker further with the representation of distinguishing keywords.

## Conclusion and Looking Ahead

Our survey of the similarity algorithms (1) Jaccard Similarity, (2) Cosine Similarity, (3) Traditional TF-IDF, (4) Word-Mover's Distance, and (5) Sent2Vec, revealed their strengths for producing reasonably accurate matches between portfolios, i.e. VCs, and individual startups, i.e. child companies. Given that we primarily sought to identify useful methods for helping entrepreneurs overcome the difficulty of information overload in their quest for funding, our results indicate that the most promising next steps would place an emphasis on accuracy in this process, possibly by using a tight threshold and less ML-oriented techniques. However, the lack of research regarding the utility of predictions produced by these ML-focused models, which were somewhat less accurate, in terms of matching startups to the portfolios that they belong to in real life, prevents us from reaching a more definite conclusion.

In other words, our findings suggest that (1) further research regarding the quality of ML-driven predictions, as well as (2) the fine-tuning, and possibly combination, of the more traditional and "accurate" NLP techniques, serve as the most promising next steps. More concretely, this might be in the form of a novel search engine for Venture Capital firms, or alternatively, the creation of an ensemble system integrating all of these surveyed methods. Additionally, in regards to the challenges posed by information restriction, the release of a curated, non-commercial dataset containing descriptions of over 30k+ companies may serve an important role in helping entrepreneurs overcome potential dataset inaccessibility obstacles.

Nevertheless, this research marks an important step forward in our quest towards unlocking a more innovative world, by helping entrepreneurs overcome initial roadblocks that often prove terminal (Bryant) for startups.

## References


The year in charts: VC defies 2020 expectations despite the pandemic | PitchBook.

Jaime Baldeon Suarez, Paloma Martínez, and Jose Luis Martínez. 2020. Combining financial word embeddings and knowledge-based features for financial text summarization UC3M-MC System at FNS-2020. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 112–117, Barcelona, Spain (Online). COLING.

Sean Bryant. How Many Startups Fail and Why?

Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. 2018. Document Similarity for Texts of Varying Lengths via Hidden Topics. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2341–2351, Melbourne, Australia. Association for Computational Linguistics.

Loet Leydesdorff, Lutz Bornmann, Rüdiger Mutz, and Tobias Opthof. 2011. Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, 62(7):1370–1381.

Xiaofeng Zhu, Diego Klabjan, and Patrick Bless. 2017. Semantic Document Distance Measures and Unsupervised Document Revision Detection. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 947–956, Taipei, Taiwan. Asian Federation of Natural Language Processing.