MGT 687, SPRING 2015

Trends in Entrepreneurship: Insights from Big Data

Aman Madaan Madaan[1*] and Prof. Shivganesh Bhargava[2]

[*]Correspondence: amanmadaan@cse.iitb.ac.in
[1]Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai, India
Full list of author information is available at the end of the article

Abstract

We explore the use of analytics on large datasets to discover various trends in Entrepreneurship. We primarily focus on 3 different kinds of trends: The rise in popularity of Entrepreneurship in the last few decades, characteristics of Entrepreneurs and the location trends. In the process, we quantitatively back several accepted, but hitherto subjective notions, and make several interesting observations.

The datasets used include the Google n-gram corpus, a dataset of 74 million tweets, Pan 14 author profiling dataset and Google trends.

Keywords: Entrepreneurship; Profiling; Data Analytics

1 Introduction

1.1 Big Data

The recent years have seen a never before surge in the amount of data that is generated around us. With the number of users having access to the internet increasing day by day, a data revolution is upon us. The exponential number of users online also leads to a large amount of opinion based data generated, especially via the social media route. Though there is no fixed definition of Big Data, IBM defines [1] big data as a dataset that is characterized by the 3Vs:

- **Volume:** 500 million Tweets are sent per day
- **Variety:** Text, videos, digitized books
- **Velocity:** Visa Inc. is capable of handling 24,000 credit card transactions per second

Since our interest lies in unearthing of trends, we will be focusing on datasets that are large and multifaceted i.e. Big Data that is *big* in either Volume or Variety.

1.2 The importance of Social Media

Social media [2] is a mirror of what is happening the world at the moment. In some sense, it captures the spirit of the masses. In the past, social media has played an important role in world changing events such as the Arab Spring [3]. Several researches have shown that there is a high correlation between what people do in real life and what they tweet about, the tweets reflect their mindset. For example, a recent study shows that analyzing tweets shows that people living closer to grocery stores make healthier food choices [4]. Entrepreneurship is a mindset, and thus we hypothesize that there would be a huge overlap between people who tweet about Entrepreneurship, and those who have an Entrepreneurial mindset, even if they are not directly involved.

1.3 Structure

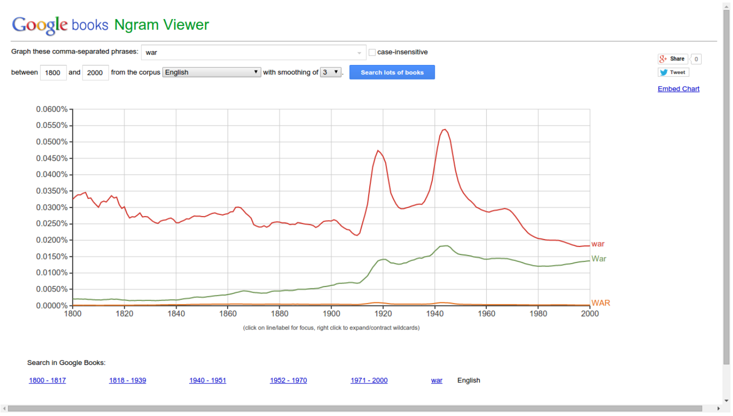
Section 2 presents some trends on Entrepreneurship obtained from the Google n-gram dataset. In Section 3 and 4, we use social media datasets to profile people interested in Entrepreneurship. Section 5 focuses on location based trends. We conclude the paper in section 6.

2 Quantitative analysis of entrepreneurship trends using millions of digitized books

2.1 Google N-gram

The Google n-gram corpus is a digital repository of 5 million books in English language printed between 1800 to 2000. This number makes up for about 4% of the books ever published. This paves way for an excellent tool that can be used to answer questions surrounding cultural evolution [5].

We motivate the use of the n-gram tool by means of 2 examples. The first example shows how the appearance of the term “war” in the books coincides with the two World wars:



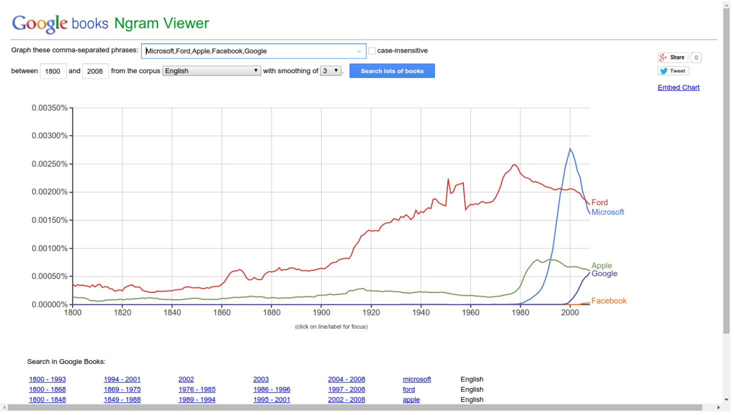
Graph these comma-separated phrases: war ☐ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Figure 1 Increase in occurrence of “war” in books: Note the peaks around the two world wars

The second figure 2.1 shows that the appearance of various companies. As can be noted, the time at which the names started appearing in the books co-incides with the time when these companies started becoming *trendy*.

Both these examples convey the fact that the books capture the Zeitgeist of the times, and are a good indicator of *what's happening*.



Google books Ngram Viewer

Graph these comma-separated phrases: Microsoft,Ford,Apple,Facebook,Google □ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 3. Search lots of books

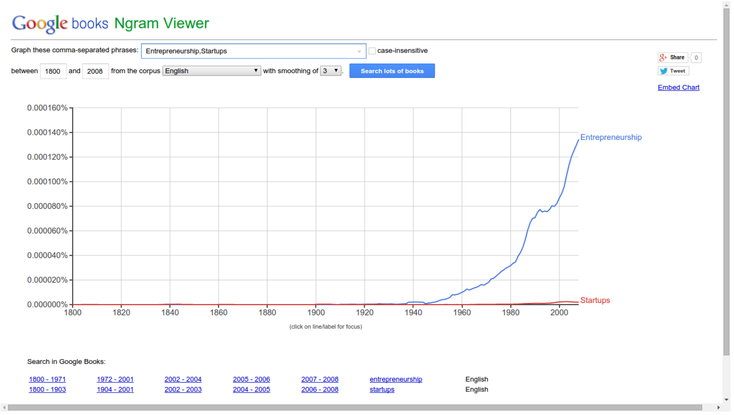
Search in Google Books:

1800 - 1993	1994 - 2001	2002	2003	2004 - 2008	microsoft	English
1800 - 1868	1869 - 1975	1976 - 1985	1986 - 1996	1997 - 2008	ford	English
1800 - 1848	1849 - 1988	1989 - 1994	1995 - 2001	2002 - 2008	apple	English

Figure 2 Various Companies as they Appeared

2.2 Entrepreneurship trends from books

2.2.1 "Entrepreneurship" vs. "Startups"



Google books Ngram Viewer

Graph these comma-separated phrases: Entrepreneurship,Startups □ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 3. Search lots of books

Search in Google Books:

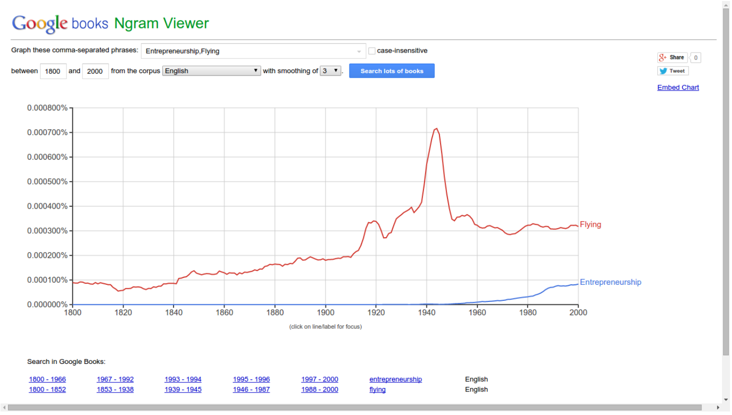
1800 - 1971	1972 - 2001	2002 - 2004	2005 - 2006	2007 - 2008	entrepreneurship	English
1800 - 1903	1904 - 2001	2002 - 2003	2004 - 2005	2006 - 2008	startups	English

Figure 3 The word "Entrepreneurship" in Books

As figure 3 shows, the term Entrepreneurship started appearing in the books around 1940, and has only evolved since then. Compare this with the term *startup*, which started figuring out in the text around 1995, and has gained traction since then. However, books still seem to be preferring the term "Entrepreneurship" over "Startup".

2.2.2 What is more risky: Entrepreneurship or Flying?

We next plot the frequency of phrases in books where the word "risk" has modified the word "Entrepreneurship". As figure 5 shows, the phrases where entrepreneurship has been modified by risk outnumber the phrases where flying has been modified by risk! This is even after figure 4 shows that flying has a higher frequency than Entrepreneurship. Again, the term "startups" start appearing around 1995, and achieve a lower frequency because of a low appearance rate as seen in figure 3.



Google books Ngram Viewer

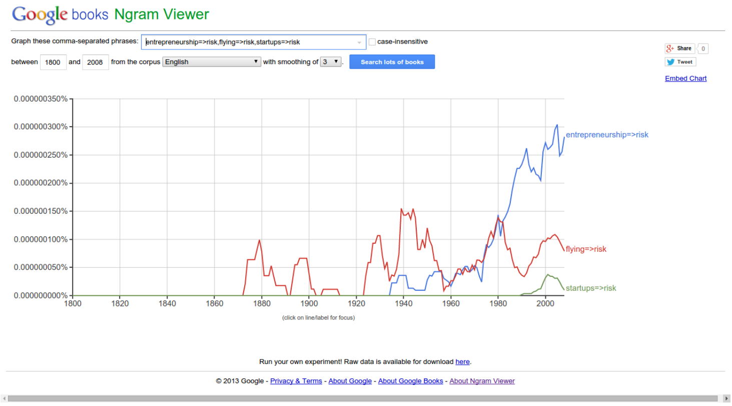
Graph these comma-separated phrases: Entrepreneurship,Flying □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Search in Google Books:

1800 - 1966	1967 - 1992	1993 - 1994	1995 - 1996	1997 - 2000	entrepreneurship	English
1800 - 1852	1853 - 1938	1939 - 1945	1946 - 1987	1988 - 2000	flying	English

Figure 4 A comparison of frequency of the terms "Entrepreneurship" and "Flying"



Google books Ngram Viewer

Graph these comma-separated phrases: entrepreneurship=>risk,flying=>risk,startups=>risk □ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 3.

Run your own experiment! Raw data is available for download here.

© 2013 Google - Privacy & Terms - About Google - About Google Books - About Ngram Viewer

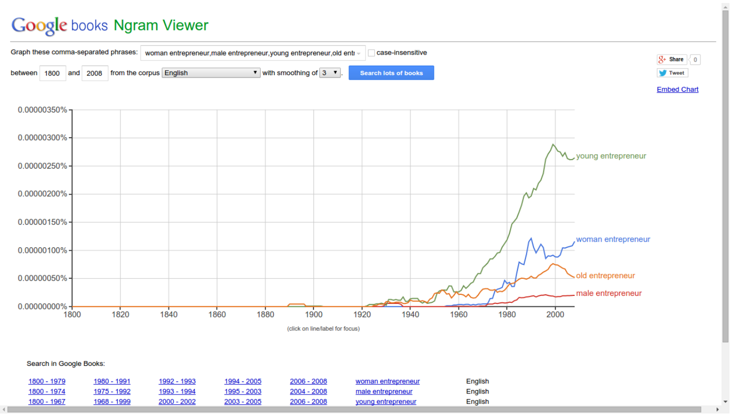
Figure 5 Entrepreneurs as risk takers

2.2.3 Young Vs. Old

We also plot "young entrepreneur", "old entrepreneur", "male entrepreneur" and "women entrepreneur". As figure 6 shows, books talk about "young entrepreneur" more than "old entrepreneur", which is somewhat expected. The high frequency of the phrase "woman entrepreneur" can perhaps be explained by noting that there has been a rise in gender equality movements around the globe in the recent times. The term "male entrepreneur" is quite unnatural and uncommon, which is likely the reason for its low frequency.

2.2.4 Entrepreneurship as a career

We next search the phrase "Entrepreneurship as a *". This is a wildcard search, and matches the top 10 phrases of the form "Entrepreneurship as a X", where X is a word. As figure 7 shows, Entrepreneurship as a field and Entrepreneurship as a career have become the trends in the recent years; it is no longer considered an esoteric career within the grasp of a select few adventurers.



Google books Ngram Viewer

Graph these comma-separated phrases: woman entrepreneur,male entrepreneur,young entrepreneur,old ent□ case-insensitive

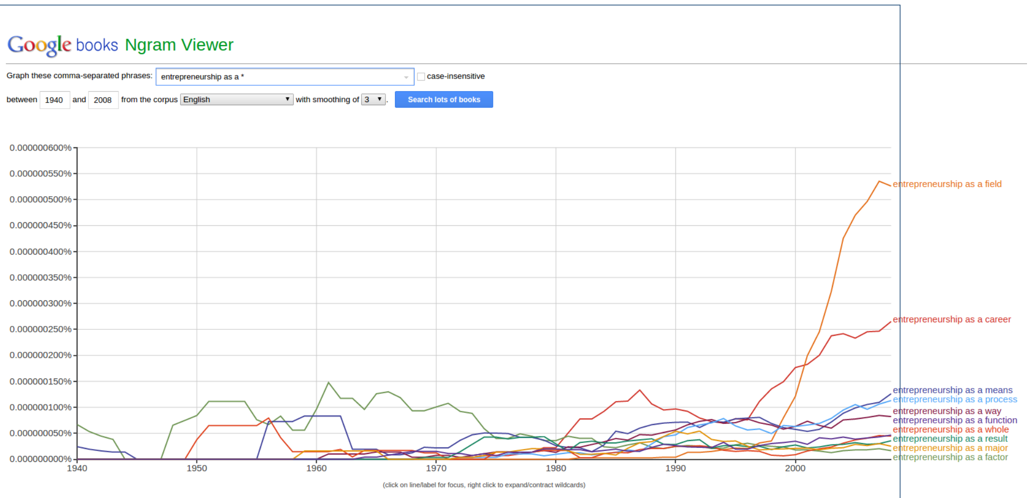
between 1800 and 2008 from the corpus English with smoothing of 3 .

(click on line/label for focus)

Search in Google Books:

1800 - 1979	1980 - 1991	1992 - 1993	1994 - 2005	2006 - 2008	woman entrepreneur	English
1800 - 1974	1975 - 1992	1993 - 1994	1995 - 2003	2004 - 2008	male entrepreneur	English
1800 - 1967	1968 - 1999	2000 - 2002	2003 - 2005	2006 - 2008	young entrepreneur	English

Figure 6 The Stereotypes carry to the books



Google books Ngram Viewer

Graph these comma-separated phrases: entrepreneurship as a * □ case-insensitive

between 1940 and 2008 from the corpus English with smoothing of 3 .

(click on line/label for focus, right click to expand/contract wildcards)

Search lots of books
0.000000600%
0.000000550%
0.000000500%
0.000000450%
0.000000400%
0.000000350%
0.000000300%
0.000000250%
0.000000200%
0.000000150%
0.000000100%
0.000000050%
0.000000000%
1940
1950
1960
1970
1980
1990
2000
entrepreneurship as a field
entrepreneurship as a career
entrepreneurship as a means
entrepreneurship as a process
entrepreneurship as a way
entrepreneurship as a function
entrepreneurship as a whole
entrepreneurship as a result
entrepreneurship as a major
entrepreneurship as a factor

Figure 7 Fraction of Phrases having "Entrepreneurship as a Career" in books from 1940 to 2008 as opposed to other phrases with "Entrepreneurship as a *"

3 Mining Tweets for Profiling Entrepreneurs

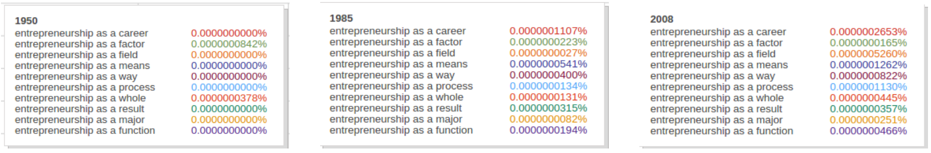
We use the Twitter N-gram corpus as provided by [6]. The dataset is annotated with gender information, as well as information about the days on which the tweets were created.

3.1 Description

As the following table from [6] shows, the dataset is almost equally divided among tweets from males and females, with *a bias towards females.*

	Female	Male	Unknown	Total
Users	3,291,849	3,062,113	5,080,426	11,434388
Tweets	21,179,637	17,190,618	36,209,710	74,579,965
Tokens	337,225,305	268,143,434	569,226268	1,174,595,007

Table 1 Gender wise tweet distribution



1950	
entrepreneurship as a career	0.0000000000%
entrepreneurship as a factor	0.0000000842%
entrepreneurship as a field	0.0000000000%
entrepreneurship as a means	0.0000000000%
entrepreneurship as a way	0.0000000000%
entrepreneurship as a process	0.0000000000%
entrepreneurship as a whole	0.0000000378%
entrepreneurship as a result	0.0000000000%
entrepreneurship as a major	0.0000000000%
entrepreneurship as a function	0.0000000000%

1985	
entrepreneurship as a career	0.0000001107%
entrepreneurship as a factor	0.0000000223%
entrepreneurship as a field	0.0000000027%
entrepreneurship as a means	0.0000000541%
entrepreneurship as a way	0.0000000400%
entrepreneurship as a process	0.0000000134%
entrepreneurship as a whole	0.0000000131%
entrepreneurship as a result	0.0000000315%
entrepreneurship as a major	0.0000000082%
entrepreneurship as a function	0.0000000194%

2008	
entrepreneurship as a career	0.0000002653%
entrepreneurship as a factor	0.0000000165%
entrepreneurship as a field	0.0000005260%
entrepreneurship as a means	0.0000001262%
entrepreneurship as a way	0.0000000822%
entrepreneurship as a process	0.0000001130%
entrepreneurship as a whole	0.0000000445%
entrepreneurship as a result	0.0000000357%
entrepreneurship as a major	0.0000000251%
entrepreneurship as a function	0.0000000466%

Figure 8 Fraction of Phrases having "Entrepreneurship as a Career" in books from 1940 to 2008 as opposed to other phrases with "Entrepreneurship as a *"

We used the 1-gram variant of the dataset for out experiments because we were only interested in mentions of entrepreneurship related buzzwords, and not word sequences that are related to entrepreneurship.

There are 4, 874, 312 words in the dataset. Each row of the dataset has information for one word, say "glass". The information includes the number of unique users who tweeted using the word "glass" in their tweets, on what days, how many of them were males and so on.

3.2 Questions Asked

We want to answer the following two questions using this dataset:

- Is one of the genders clearly more interested in entrepreneurship from a social media perspective?
- Do people think about startups/entrepreneurship more on some days than the others?

3.3 Methodology

We zeroed in on several buzzwords that indicate activities related to entrepreneurship. Primarily 4 categories were used:

- Entrepreneurship
- Startups
- Venture Capitalist
- Angel Investors
- Founder

We were able to find 268 relevant keywords. Examples include *@startuppro*, *entrepreneurialism* and *@freshfounder*. A complete list of all the keywords can be located in table 1 in the appendix.

3.4 Results

Is one of the genders clearly more interested in entrepreneurship from a social media perspective?

From our study of the dataset, it becomes pretty clear that **males are much more involved** in tweets pertaining to entrepreneurship and related fields.

Day specific pattern of tweeting We observe high activities on weekdays, and relatively low activities on weekends.

3.4.1 Sanity Check

To drive our point and as a verification of our methodology and datasets, we decided to find user proportions for topics that are stereotypically "female" and stereotypically "male".

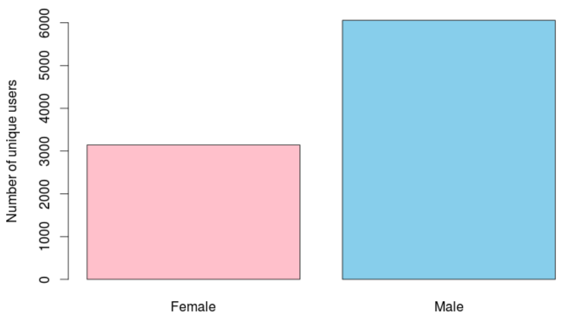


Figure 9 Gender Distribution of Users with tweets having Entrepreneurship related buzzwords

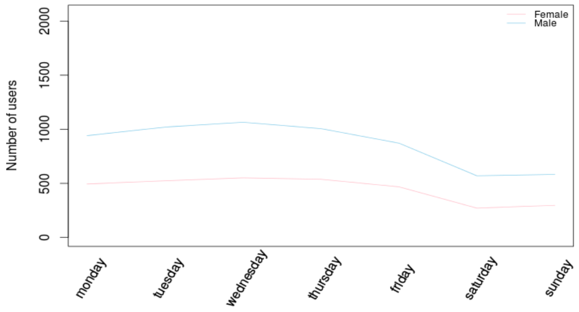


Figure 10 Per day Distribution of Male and female Users with tweets having Entrepreneurship related buzzwords

For females, we chose the topic "dress" and for males, we chose "cricket". As can be seen from the following plots, females dominate the topic "dress" and males dominate the topic "cricket". This thus strengthens our hypothesis that Entrepreneurship attracts more male attention than female attention.

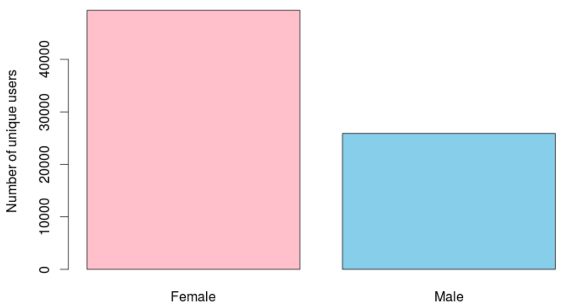


Figure 11 Gender Distribution of Users with tweets having "dresses" related buzzwords

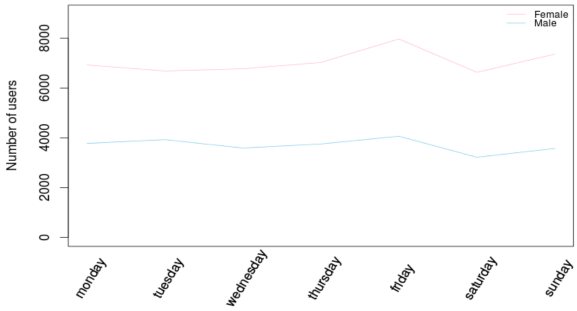


Figure 12 Per day Distribution of Male and female Users with tweets having "dresses" related buzzwords

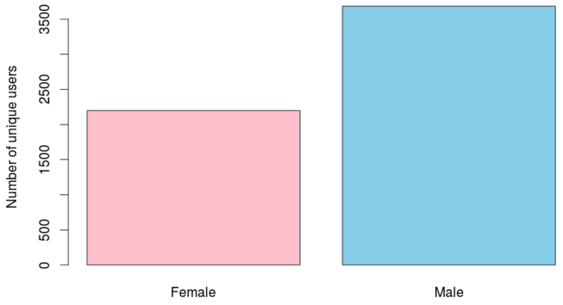


Figure 13 Gender Distribution of Users with tweets having "cricket" related buzzwords

4 Age Distribution of Users Blogging about Entrepreneurship

4.1 Dataset

The PAN 2014 dataset [7] provides tweets and blogs annotated with the age and gender of the user. There were not sufficient tweets about entrepreneurship in

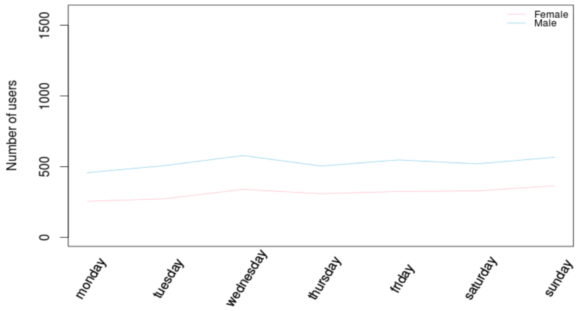


Figure 14 Per day Distribution of Male and female Users with tweets having "cricket" related buzzwords

the dataset. However, there were 1105 unique users who had written about entrepreneurship on the social media.

Their age and gender distribution is discussed in the next subsection.

4.2 Analysis

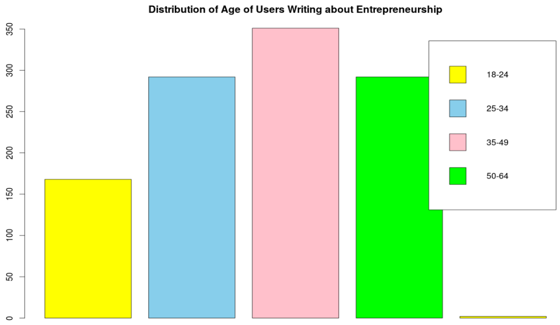


Figure 15 Age Distribution of Users Writing about Entrepreneurship on the social media from the PAN 14 dataset

Age Group	Count
18-24	168
25-34	292
35-49	351
50-64	292
64-	2

Table 2 Age Distribution of Users Writing about Entrepreneurship on the social media from the PAN 14 dataset

The distribution seems to be forming the classical normal curve. However, as Table 2 shows, almost a third of the users are below the age of 34. It may be noted that the original dataset had equal number of male and female bloggers.

Gender	Count
Female	520
Male	585

Table 3 Gender Distribution of the users writing about Entrepreneurship the PAN dataset

Table 3 shows the distribution of gender. Though the distribution is not as skewed as the twitter dataset, males again outnumber females.

5 Location Based Trends from Google Search Logs

Google is the most popular search engine used these days. According to different surveys, the user percent of Google stands at around 70% [8]. With such a popular user base, it may be expected that Google search logs can give a sense of what is popular around the globe these days. In fact, Researchers have explored the possiblity of using the search logs to determine if there is a possiblity of an epidemic. The idea being that a large number of searches related to flu symptoms may indicate that a large number of people are suffering from flu and thus may be a signal of an impending epidemic [9] .

The idea of this section is to explore the Google search trends for terms that indicate entrepreneurship at and around locations that are traditionally thought of as "startup" hubs. The underlying assumption is that if people are either involved in or are enthusiastic about entrepreneurship, they would look it up on the Internet, most likely using Google [8].

5.1 Questions Asked

The questions we want to ask in this section are the following:

1 What are the different locations where entrepreneurship forms a topic of interest for the masses?

2 Are the locations obtained in 1) the places that are traditionally considered to be "startup hubs"? Is there some pattern of geographical influence flow that is apparent?

3 Is the term "startup" more popular than the term "Entrepreneurship" in some parts of the world and vice-versa?

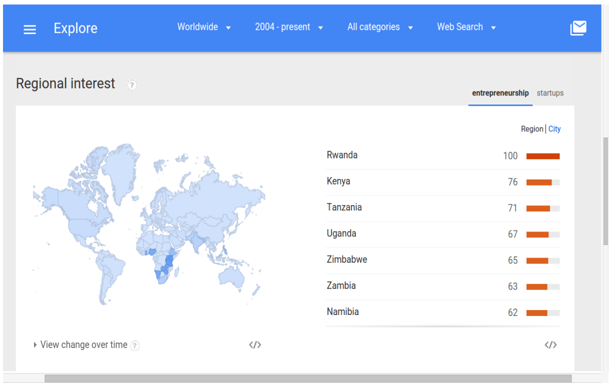
5.2 Results

5.3 Conclusions

5.3.1 What are the different locations where entrepreneurship forms a topic of interest for the masses?

The top countries for the keyword "entrepreneurship" are all part of the African Subcontinent, with Rwanda, Kenya taking the top spot. This should not be surprising, since a recent survey [10] has revealed that out of the top 10 countries where Adults in the age group of 18-64 are in the early stage entrepreneurship activity, top 4 are from African Subcontinent, with Ghana at the fourth place with 25.8%! Another survey shows that 71.5% of the Africans see Entrepreneurship as a good career choice [11].

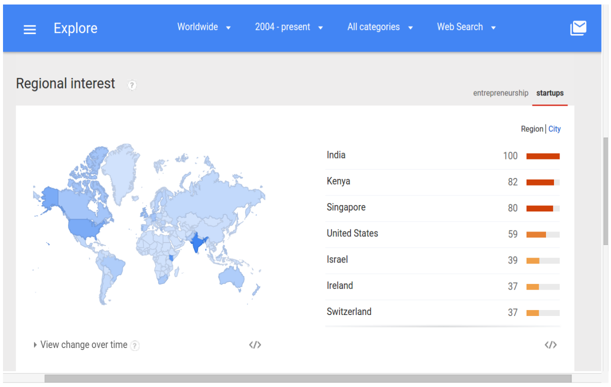
For the term startup, the results are more predictable, with India dominating the list, followed by the expected names like Singapore and US. City wise trends also paint a predictable picture.



Worldwide · 2004 - present · All categories · Web Search

Region	
Rwanda	100
Kenya	76
Tanzania	71
Uganda	67
Zimbabwe	65
Zambia	63
Namibia	62

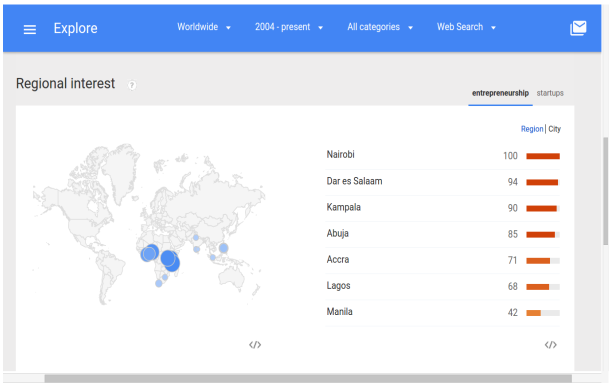
Figure 16 Regional Interest in the term “entrepreneurship” for 2004-present



Worldwide · 2004 - present · All categories · Web Search

Region	
India	100
Kenya	82
Singapore	80
United States	59
Israel	39
Ireland	37
Switzerland	37

Figure 17 Regional Interest in the term “startups” for 2004-present



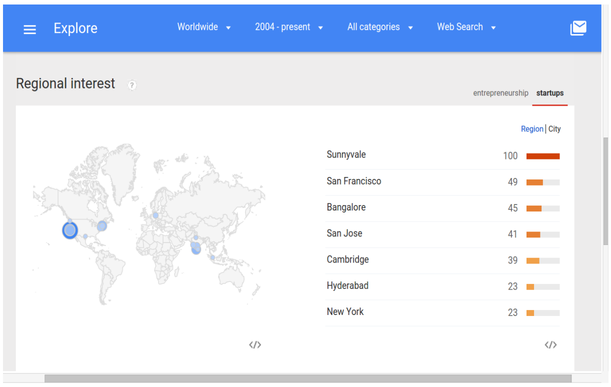
Worldwide · 2004 - present · All categories · Web Search

City	
Nairobi	100
Dar es Salaam	94
Kampala	90
Abuja	85
Accra	71
Lagos	68
Manila	42

Figure 18 City wise Interest in the term “entrepreneurship” for 2004-present

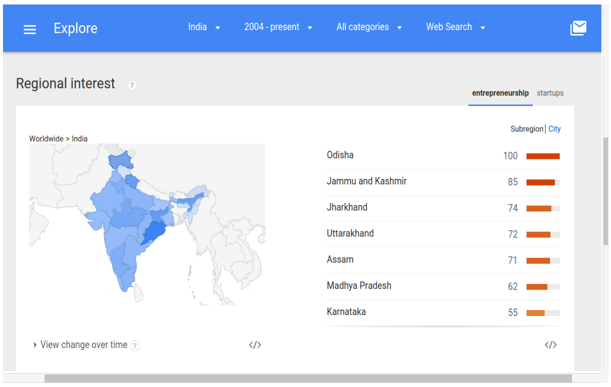
5.3.2 Are the locations obtained in 1) the places that are traditionally considered to be “startup hubs”?

Yes, especially if you look at the trends for the term “Startups”. Sillicon Valley is known around the globe as the starting point of a large number of startups, and



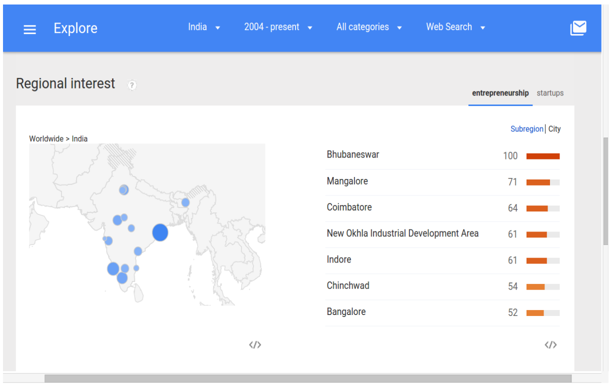
Sunnyvale	100
San Francisco	49
Bangalore	45
San Jose	41
Cambridge	39
Hyderabad	23
New York	23

Figure 19 City wise Interest in the term "startups" for 2004-present



Odisha	100
Jammu and Kashmir	85
Jharkhand	74
Uttarakhand	72
Assam	71
Madhya Pradesh	62
Karnataka	55

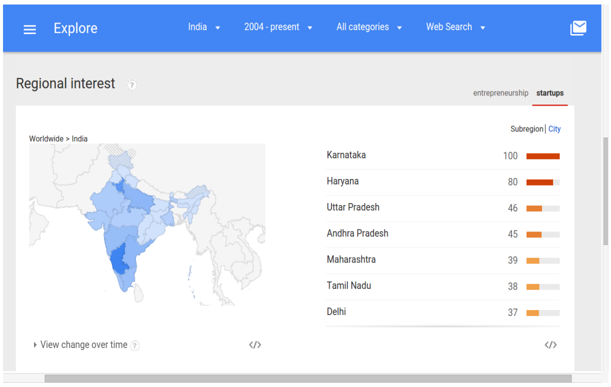
Figure 20 Sub-Regional Interest in the term "entrepreneurship" for 2004-present in India



Bhubaneswar	100
Mangalore	71
Coimbatore	64
New Okhla Industrial Development Area	61
Indore	61
Chinchwad	54
Bangalore	52

Figure 21 City Interest in the term "entrepreneurship" for 2004-present in India

tops the list. Indian startup hubs like Bangalore and Hyderabad also appear on the list.



Explore India 2004 - present All categories Web Search

Regional interest

entrepreneurship startups

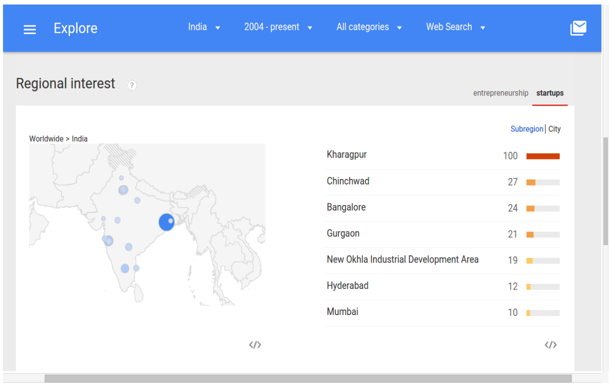
Worldwide > India

Subregion | City

Karnataka	100
Haryana	80
Uttar Pradesh	46
Andhra Pradesh	45
Maharashtra	39
Tamil Nadu	38
Delhi	37

View change over time

Figure 22 Sub-RegionalInterest in the term "startups" for 2004-present in India



Explore India 2004 - present All categories Web Search

Regional interest

entrepreneurship startups

Worldwide > India

Subregion | City

Kharagpur	100
Chinchwad	27
Bangalore	24
Gurgaon	21
New Okhla Industrial Development Area	19
Hyderabad	12
Mumbai	10

Figure 23 City Interest in the term "startups" for 2004-present in India

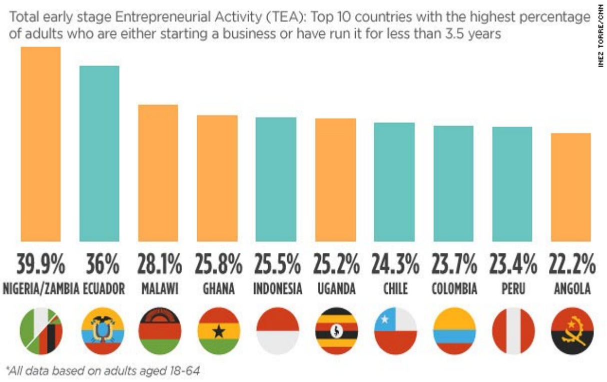


Figure 24 Reproduced from [10]

5.3.3 Is there a pattern of geographical influence flow that is apparent?

To an extent, yes. It is clear from the Indian sub-region pattern that states that host cities where entrepreneurship is popular tend to be the ones from where highest number of queries flow in. All the expected names appear in the list. Kharagpur forms an outlier, but that is most likely because of the queries from IIT Kgp.

5.3.4 the term "startup" more popular than the term "Entrepreneurship" in some parts of the world and vice-versa?

Yes, considering the fact that most of the startups in the Sillicon Valley and in Bangalore in India are tech based, it can be said that in **the tech based new venture world, the term startup is more common**.

6 Conclusions

- Entrepreneurship started appearing in books around 1940s, and has since then only gained popularity. Entrepreneurs are mostly young (books mention "young entrepreneurs" more than "old entrepreneurs"), and Entrepreneurship is considered risky (risk associated with entrepreneurship is more talked about than the risk associated with flying).
- Entrepreneurship started getting popular as a career, and as a field around 1995, when the world saw the rise of the Microsofts and Apples.
- Usually, males are more interested in Entrepreneurship than females, and are most likely working (high number of tweets on the weekdays and a dip on weekends). Most of them are below the Age of 35.
- African Subcontinent is fast becoming an Entrepreneurship hub. The classical centers (Sillicon Valley in the US, Hyderabad, Bangalore, Gurgaon, Noida in India) still witness a lot of buzz around Entrepreneurship, but there is no concentration of activity. There is considerable interest from all around the globe.

Acknowledgements

Discussions with Aanchal Parker were extremely helpful in getting an overall picture of entrepreneurs in present times. Thanks to the class of MG 687 for insightful discussions, many of which have improved our understanding of entrepreneurship.

Author details
[1]Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai, India.
[2]Shailesh J. Mehta School of Management , Indian Institute of Technology Bombay, Mumbai, India.

References

1. IBM Big Data, Http://www-01.ibm.com/software/in/data/bigdata/
2. Kaplan, A.M., Haenlein, M.: Users of the world, unite! the challenges and opportunities of social media. Business horizons **53**(1), 59–68 (2010)
3. Howard, P.N., Duffy, A., Freelon, D., Hussain, M., Mari, W., Mazaid, M.: Opening closed regimes: what was the role of social media during the arab spring? (2011)
4. Chen, X.: Analyzing tweets shows that people living closer to grocery stores make healthier food choices. LSE American Politics and Policy (2014)
5. Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., *et al.*: Quantitative analysis of culture using millions of digitized books. science **331**(6014), 176–182 (2011)
6. Herdağdelen, A.: Twitter n-gram corpus with demographic metadata. Language resources and evaluation **47**(4), 1127–1147 (2013)
7. The PAN 14 Dataset, Http://pan.webis.de/
8. Who Is Winning the Search Engine War?, Http://searchengineland.com/whos-really-winning-search-war-204651
9. Dugas, A.F., Hsieh, Y.-H., Levin, S.R., Pines, J.M., Mareiniss, D.P., Mohareb, A., Gaydos, C.A., Perl, T.M., Rothman, R.E.: Google flu trends: correlation with emergency department influenza rates and crowding metrics. Clinical infectious diseases **54**(4), 463–469 (2012)
10. Africa Is Buzzing with Entrepreneurial Spirit, Http://edition.cnn.com/2014/05/13/business/numbers-showing-africa-entrepreneurial-spirit/
11. 71.5% of the Africans See Entrepreneurship as a Good Choice, Http://disrupt-africa.com/2015/02/71-5-africans-see-entrepreneurship-good-career-choice/