Identifying Topics about Leadership and Entrepreneurship using Topic Modelling

Silvia Beatriz González-Brambila, Josué Figueroa-González

Universidad Autónoma Metropolitana, Distrito Federal, Mexico

{sgb,jfgo}@correo.azc.uam.mx

Abstract. Terms like leadership and entrepreneurship have reached great importance and popularity. A lot of information has appeared related with these topics and searching for it has become a bit complicated for people interested in leadership and entrepreneurship. With so much information, it is difficult to identify a topic among the huge quantity of concepts involved in these themes. This work presents a process based on text processing, text mining and topic modelling for finding topics and patterns in interviews about these topics in order to identify the main themes that leaders and entrepreneurs talk about.

Keywords: Topic model, text mining, knowledge discover, leadership, entrepreneurship.

1 Introduction

Concepts like leadership and entrepreneurship have grown in the last years, nowadays there are research branches, schools of thoughts and a lot of information like books, speeches, conferences, workshops, etc. has appeared. Even, researches about characteristics that a person must have in order to be considered a leader or an entrepreneur have been done [9]. Many people are interested in these topics and look up for any information that allows them to become a leader or an entrepreneur.

There are a lot of topics related with these terms, for example: concepts that a leader or an entrepreneur must know, suggested books and authors for reading, conferences, seminars, and leaders or entrepreneurs experiences and talks. Internet has become one of the main sources of information for almost any topic, in any format. However, the amount of information makes difficult its search and its easy to be overwhelmed with a lot of concepts and ideas. Talks, conferences, speeches and interviews appear every day in different formats, as documents, web pages, video and audio. People like listening experts and well know leaders and entrepreneurs, but even this subset of information contains a lot of terms, concepts and ideas, so it becomes difficult to identify the main topics, especially when a person its not related with these two popular themes.

For this reason, a research over leadership and entrepreneurship is proposed using concepts related with text mining and topic models. The information considered for the work is contained in audio files, common known as podcast, a media that has grown considerably during the last years thanks to the Internet. The main contribution of this works is the analysis of concepts like leadership and entrepreneurship obtained from interviews with leaders and entrepreneurs using the Topic Model framework in order to identify what do they talk about.

This paper is composed by the following sections: section 2 presents the basic ideas of the elements and frameworks involved in the process of text analysis. Section 3 presents the description of the steps performed during the analysis considering as base Text Mining and Topic Modelling processes. Section 4 contains the results obtained in each step and their interpretation. Finally, section 5 contains the conclusions and future research.

2 Knowledge Discovery

Several techniques have appeared in order to manage, process, discover, understand and exploit knowledge from any kind of information and format. Many of these concepts have similar processes and performance, so it could be a little complicated distinguish one from another, for example: Knowledge Discover on Databases, Topic Modelling, and Data and Text Mining.

All of them can be used for data analysis and information retrieval which give a certain value to the users. However, even if their general process can look similar, there are specific differences that distinguish each other.

2.1 Knowledge Discover on Databases

Process for extracting useful information has received many names, for example: Data Mining, Text Mining, Knowledge Extraction or Information Discover, but nowadays these concepts are being considered as part of Knowledge Discover on Databases (KDD).

There are several definitions for KDD [7], but in general it can be resumed as the process for obtaining or discovering knowledge from data using different processes, for example: Data Mining which is considered as a step in the whole KDD process.

2.2 Data Mining

Data Mining is the process of applying specific algorithms for obtaining patterns in data [8]. Its main objectives can be classified in: verification and discovering.

In verification, the system only verifies a hypothesis, meanwhile in discovering, the goal is to automatically find new patterns in data. Applied, these two goals are transformed in prediction and description.

In order to reach these goals, Data Mining uses a set of techniques like: classification, regression, summarizing, dependencies modelling and other that

can be grouped in: statistical for hypothesis validation or learning methods for recognizing patterns, however some of the techniques can be shared between both groups [5].

2.3 Text Mining

Text Mining consists in obtaining patterns of interest from documents or data, especially when these patterns are not so obvious even for experts [13]. The main difference with Data Mining, is that Text Mining works with non-structured data, for example, words in a document, and Data Mining works with structured data, like the stored in Data Bases [14].

Text Mining can be seen as an option for Data Mining for a special kind of data, it also can be used like a process similar to KDD, also, Text Mining can be used as a tool for structuring data contained in a group of documents so them can be analyzed with Data Mining techniques.

2.4 Topic Modelling

Topic Model [13] is a mathematical framework which main purpose is to discover the hidden patterns in a document in order to identify the topics that are in a group of documents for exploiting their knowledge. It assumes that in any document, some words will appear more frequently than other, and using some statistical analysis of these words, identifies the different topics that can be found in a group of documents, identifying the relationship between them.

This technique is used specially over non structured data and documents collections, for example: texts, web sites content, comments over social networks [11, 10] and e-mails. Topic Models algorithms can be used also for analyze genetic information and images.

Topic model algorithms can be classified into two main groups: sampling-based and variational, the most common algorithms are: Gibbs [4] for sampling and Variational Estimation Maximization (VEM) [15] for variational ones.

2.5 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [3] is a hierarchical Bayesian Model used for obtaining statistical relationships between groups of documents. LDA is based in two concepts:

- Any document is composed by a distribution of topics,
- Each topic is a distribution of words.

LDA assumes that a topic is a distribution over a fixed set of words, and those topics have been specified before data were generated. Then, the generation of a document is performed in the following way:

- A topic distribution is chosen,

– For each word in a document:
  • A topic from the topic distribution is chosen randomly,
  • A word is randomly chosen from the vocabulary distribution of the previous selected topic.

For LDA, a group of documents share the same topics, but each document presents them in different proportions. Documents are observable data; meanwhile the structure of topics, their vocabulary and distribution per document are the hidden data. LDA looks up for discovering the hidden data or structures using an inverse process to the one that generates the document.

3 Discovering Topics

Although the work of this paper was focused in analyze the content of audio files, podcast, their information was extracted to text files, so, the process for analyzing the information and finding the most common terms about leadership and entrepreneurship is related to Text Mining and Topic Model.

For the information analysis, it was used the software R, especially the libraries related with text processing and topic modelling: *tm*, *slam*, *topicmodels*, *snowballC*, and *clue*.

3.1 Obtaining and Integrating Information

A group of podcast related with leadership and entrepreneurship were obtained from different sources [1], also some transcription from interviews were considered [12, 6]. From the amount of available interviews, the selected ones talked about different topics, for example, products or business, and concepts about becoming and being a leader or an entrepreneur.

In the Integration step, the content of the audio was expressed as a text file using the application Adobe Premier Pro CC. Also, other text files were generated from the content of the transcription from the interviews. The analysis was performed with different kind of information: interviews with leaderships, with entrepreneurs and combining some interviews of both types.

The fact of choosing interviews which combine general concepts of leadership and entrepreneurship with concepts about different products and business, offered an interesting analysis of the information.

3.2 Selection, Cleaning and Transformation

This process, also known as data cooking, processes the data in order to obtain a more appropriate representation for their analysis. For the content obtained from audio files, punctuation marks didn't exist, but other elements which could be easy removed with some routines in R, like numbers, here appeared as words, so a special process had to be applied to eliminate them.

The *Stop Words*, which are words considered as the most common in a language, had to be filtered before the topic classification process. The same

occurs with blank spaces and other group of words (pronouns, connectors and articles). A word can be reduced to its original root, this is called stream process, some tests were performed with steamed documents and other without applying this step.

The format of the interviews obtained from web content contains the names of the interviewer and the interviewed, after some tests, these names appeared as frequent words in many topics, so them had to be cleared from the text files. The amount of topics was tested from 2 to 20, obtaining similar results; finally, the decided quantity of topics chosen for the analysis was 8.

3.3 Topic Modelling

After the data cooking process, Topic Modelling framework was applied. With a fixed number of topics selected, the topic identification was realized with three models: VEM, VEM with an α fixed and Gibbs. Once the topics were obtained, a review of how close were the results for each model was performed using the concepts of Hellinger distance and the Linear Sum Assignment Problem (LSAP), this analysis was performed only for VEM with VEM using an α fixed and VEM with Gibbs.

3.4 Topic Interpretation

Once the topics and the vocabulary for each one were obtained, the next step was interpreting them. To do this, were considered the most 5 frequent words in each topic, and according to them, the main topics were identified. The results are shown in Section 4.

4 Results

In this section, are presented the results obtained in each step of processing the information.

4.1 Selection, Cleaning and Transformation

After processing the podcast and the interviews, a group of files was created, one composed only for data related with leadership, another about entrepreneurship and the last one with mixed interviews.

Each group was composed by ten interviews, and for the combined group, were randomly chosen five for each one of the other groups.

After the cleaning process performed with R routines, were obtained the following quantity of terms and documents: Leadership, 1960 terms and 176 documents. Entrepreneurship, 1368 terms and 146 documents. Mixed, 1987 terms and 183 documents.

4.2 Topic Modelling

For the Topic model process, were proposed 8 topics per group of interview, thought some tests were performed with different amounts of topics obtaining similar results. The Hellinger distance for entrepreneurship documents between terms identified using VEM and VEM with α Fixed was 0.27678, and with VEM and Gibbs was 0.74399. For leadership using VEM and VEM with α Fixed was 0.24620, and with VEM and Gibbs was 0.75104. Finally, the distance for mixed documents using VEM and VEM with α Fixed was 0.26176, and with VEM and Gibbs was 0.75303.

4.3 Topic Interpretation

In order to identify the topics obtained, were considered the 5 most used terms in each topic. Only the topics identified with VEM and Gibbs are presented because VEM and VEM with α Fixed models produced almost the same terms.

Some topics in the entrepreneurship interviews were related with a product or business, for this reason, these terms aren't shown in the results tables, which only present the terms and topics that talk about general concepts of entrepreneurship. These results are shown in Table 1 and Table 2.

Table 1. Relevant topics and terms identified with VEM for entrepreneurship interviews.

Topic 1	Topic 2	Topic 4	Topic 6	Topic 7	Topic 8
business	time	companies	services	service	business
legal	portfolio	business	costs	companies	costs
money	career	consulting	cost	costs	think
disposable	work	source	reduction	customers	smart
working	make	product	consulting	time	idea

Topics not considered talk more about a certain product or business containing terms like: restaurant, online, product and career, this is because some of the businesses owned by an entrepreneur were from food and online services.

Terms and topics identified with Gibbs model that are not included in the results contains terms like: online, site, software and transportation.

Topics and terms identified for entrepreneurship are very clear. Talking about this theme, some main aspects which are very important for an entrepreneur can be identified. For example:

- For starting a business, many entrepreneurs talk about studying the market and having an idea, Topics 8 in VEM and 7 in Gibbs models are related to this aspect.
- Financial aspects are very important too, Topic 1 in VEM model is related with them. Management its very important, Topic 6 in Gibbs model is related with this concept.

Table 2. Relevant topics and terms identified with Gibbs for entrepreneurship interviews.

Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
market	time	need	company	services	business
value	able	think	opportunity	costs	people
first	help	selling	products	source	think
new	sales	product	service	consulting	need
venture	better	customer	make	reduction	idea

As the results show, some topics of every model are related with more than one concept because talking about entrepreneurship involves a lot of concepts that can appear in different ideas.

Topics and terms related with leadership are more general because interviews talked about general concepts of this theme, not about a certain product like in the entrepreneurship ones, for this reason, more topics are presented in the results. Table 3 and Table 4 present the topics for leadership interviews.

Table 3. Relevant topics and terms identified with VEM for leadership interviews.

Topic 1	Topic 2	Topic 3	Topic 5	Topic 6	Topic 7
people	cash	years	design	cares	time
time	business	business	vision	leader	market
company	leadership	plan	process	patient	energy
process	people	data	organization	organization	important
organization	make	experience	people	will	world

Table 4. Relevant topics and terms identified with Gibbs for leadership interviews.

Topic 1	Topic 2	Topic 5	Topic 6	Topic 7	Topic 8
leaders	world	time	things	market	people
organization	data	work	need	change	business
will	across	years	want	global	environment
culture	well	understand	job	strategic	important
plan	much	think	patient	different	meetings

Topics identified with leadership and their respective terms, also can be grouped according main themes of this concept.

– Topic 6 of VEM and Topic 1 in Gibbs can be related with the characteristics of a good leader must have.

- Topic 2 of VEM and 7 of Gibbs are related with financial aspects, this is because most of the interviews were about leaders in an business.
- Topics 3 from VEM and 5 from Gibbs can be identified as planning and time, concepts really important for a good leader.
- Topics 1 and 5 from VEM and Topic 8 from Gibbs models are related with organization of an enterprise or business.
- Topic 7 from VEM and Topics 2 and 6 from Gibbs can be related with concepts that leaders must manage.

Terms obtained from mixed interviews show a vocabulary more related with leadership with some words referring to entrepreneurship, like customers or business. With these interviews, VEM method produced better results than Gibbs. Notice than the terms about a specific product or business disappear. Table 5 and Table 6 present the topics for mixed interviews.

Table 5. Relevant topics and terms identified with VEM for mixed interviews.

Topic 1	Topic 4	Topic 5	Topic 8
time	people	care	people
good	business	example	plan
need	process	patient	data
thing	things	leaders	talk
things	important	change	meetings

Table 6. Relevant topics and terms identified with Gibbs for mixed interviews.

Topic 1	Topic 2	Topic 4
leaders	customers	business
organization	important	world
culture	certain	plan
leadership	different	care
might	know	working

Using the Hellinger distance and LSAP framework, the topics from VEM and Gibbs with best relationship were obtained. The relationship between topics from every kind of interview and every model is presented in Table 7

5 Conclusions and Future Work

It's interesting to observe the results, where there are many terms which apparently don't have a lot in common between them; however, this is mainly because

Table 7. Main relationships between VEM and Gibbs models for leadership and entrepreneurship topics.

Interviews	VEM	Gibbs
Leadership	Topic 6	Topic 6
	Topic 6	Topic 6
Entrepreneurship	Topic 2	Topic 8

of the different products and business that entrepreneurs talked about in their interviews, something that didn't occur with the leadership interviews, where there are more topics related to the main theme.

There are some topics that clearly identify aspects related with these two themes. The fact of combining both types of interviews produced a result a little different than just merging the main terms obtained separately. It's important to notice that the main topic in the mixed interviews is related to leadership, with just some words about entrepreneurship, specially talking about products, money or business.

Talking about the steam process, applying this step reduces the amount of terms per topic, however sometimes it's difficult to understand the results viewing only a part of the word. It's more important to eliminate the most frequent words; besides the routines available in R, it's recommended to apply another cleaning process for this words.

In order to improve the analysis, more interviews could be considered. Also these interviews mustn't be related to a product or business, but concepts about becoming, being and thinking as a leader or entrepreneur. Using other kind of sources, like speeches or talks could also contribute for obtaining better results.

About the models used for the analysis, VEM and VEM Fixed, produced better results than Gibbs. VEM Fixed results aren't shown, but they are almost the same as VEM.

The most difficult part of the whole process is to identify the topics according the vocabulary contained in each one. It's necessary to rely on an expert in the topic, so the results can be interpreted in a better way.

As future works, once the main topics and terms about leadership and entrepreneurship are known, it's planned to structure them in order to perform a Data Mining analysis for knowing more specific information about a particular theme, for example:

- Evolution of topics over time,
- Books, conferences, talks or workshops that someone interested in the topics should read or attend,
- Concepts about leadership and entrepreneurship in a specific area, for example: sports, education, business, politics, etc.

Hundreds of interviews and other kind of information in many formats about almost any theme are appearing every minute over Internet. Extracting the

content of audio files makes easier the process of obtaining information, because many of the data is being generated in audio and video files.

Another topic that can be very interesting for studying is the one related with saving money, analyzing interviews, speeches, conferences from experts in the called financial education.

Processing this content with the Topic Model framework, offers a great possibility for studying and discovering knowledge from this huge amount of information that otherwise could be difficult to explore.

References

1. Audio Books, Podcasts, Videos, and Free Downloads to Learn From, http://www.bbc.co.uk/podcasts
2. BBC Podcast and Downloads, http://www.bbc.co.uk/podcasts
3. Blei, D.: Probabilistic Topic Models. Communications of the ACM. 4, 77–84 (2012)
4. Casella, G., George, E: Explaining the Gibbs Sampler. The American Statistician, 3, 167–174 (1992)
5. Chen, L., Sakaguchi, T., Frolick M.: Data Mining Methods, Applications, and Tools. Information Systems Management, 17, 65–70 (2006)
6. Entrepreneur interviews, http://www.entrepreneurship-interviews.com
7. Fayyad, U., Piatesky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine, 3, 37–53 (1996)
8. Fayyad, U., Piatesky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM, 11, 27–34 (1996)
9. Fernald, W. Jr., Solomon, G., Tarabishy.: A New Paradigm: Entrepreneurial Leadership. Southern Business Review, 2, 1–11 (2005)
10. Ghosh, D., Guha, R.: What are we tweeting about obesity? Mapping tweets with topic modelling and Geographic Information System. Cartography and geographic information science, 40, 90–102 (2013)
11. Hong, L., Davison, B. D.: Empirical study of topic modelling in twitter. In: Proceedings of the First Workshop on Social Media Analytics (2010)
12. Interviews with leaders articles and insights, http://www.mckinsey.com/insights/leading_in_the_21st_century/interviews_with_leaders
13. Mooney, R., Nahm,Y.: Text Mining with Information Extraction. Multilingualism and Electronic Language Management. In: Proceedings of the 4th International MIDP Colloquium (2005)
14. Ronen, F., et al.: Text mining at the term level. In: Principles of Data Mining and Knowledge Discovery, Springer Berlin Heidelberg, 65–73 (1998)
15. Tzikas, D. G., Likas, C. L., Galatsanos, N. P: The variational approximation for Bayesian inference. Signal Processing Magazine, 6, 131–146 (2008)