IMPACT ASSESSMENT OF "RACING EXTINCTION"

Harathi Korrapati, Jana Diesner, Rezvaneh Rezapour

FEBRUARY 2, 2016

UNIVERSITY OF ILLINOIS, GSLIS

TABLE OF CONTENTS

INDEX OF TABLES AND FIGURES

This work is supported by the FORD Foundation's JustFilms Fund and Vulcan Production. 17

1. EXECUTIVE SUMMARY

Impact assessment should begin with a research question. In the case of “Racing Extinction” (d. Louie Psihoyos, 2015), our team was asked to investigate the following questions: How has the film moved the public awareness about effects of global environmental pollution and its impact on the endangered species? And what does society think about the relation between endangered species and the environmental issues around the world?

Using key terms provided by the filmmaker, we are conducting an analysis of news media and social media discourse on the topic of the film to determine its state before the film was released. This provided the baseline for any observable change. We then looked at the discourse on the topic after the release of the film. We define impact as any observable change in the media and

social media discourse between these two points: before and after the film's release. We would like to compare and evaluate different techniques for measuring the increase in public awareness for an issue, which is an impact goal shared across most of the recent frameworks for change assessment. This basically addresses how well the content of an information from the movies has been injected in to public.

Considering the first research question, if the film had an impact on media and social media discourse, we would expect to see several things. Because we must assume that the discourse of the topic is not static (meaning it is not completely dependent on the film), we can only determine that the film (and not something else) had an impact if the media discourse on the topic of the film (extinction of animal species) intersects with the language or content of the transcript of the film in a way that it did not before the film's release. Considering the second research question, we are observing how the media discourse and social media discourse characterized the relationship between extinction of animal species and the increase of environmental pollution around the world. To do this, we are looking at the sentiments expressed, the key figures of the conversation in the content of the semantic networks before and after the release of the film.

It is important to say a brief word on data because all impact assessment is dependent on the type of data that is collected. A report is only as valid as the data is valid and appropriate. In other words, whether or not research questions can be asked or answered depends on the type of data the researcher (and filmmaker) collects and has access to. For the "Racing Extinction" film media analysis, we are analyzing the data that we collected from press coverage using the keywords for the film for a specific time period (March 23, 2014 to January 23, 2015 [= Press Before] and January 24, 2015 to November 24, 2015 [= Press After]. The split point reflects the point in time when Racing Extinction was screened at the 2015 Sundance Film Festival). These were found through LexisNexis Academic, one of the world's largest online electronic libraries for legal, business, news, and public information. For the "Racing Extinction" social media analysis, we analyzed data that we collected from Twitter. In both cases, our team used **a)** network analysis to detect key agents and organizations and **b)** text mining techniques to find trends in current discussions (topics, sentiments, dynamics).

2. INTRODUCTION

Documentaries are meant to tell a story, i.e. to create memories, imagination and sharing (Rose, 2012). Moreover, the goal with documentaries is to induce change in people's knowledge and/ or behavior (Barrett & Leddy, 2008). How can we know if a production has achieved these goals? We herein apply an empirical, scalable and systematic methodology that we have been developing for this purpose to "Racial Extension" media data.

In a nutshell, we approach this question by combining computational techniques from data mining and network analysis: we assume that documentaries are produced, screened and watched as part of larger and continuously changing ecosystems that involve multiple stakeholders and the flow of information between them. We track, map and analyze socio-semantic networks that represent these stakeholders and the information they disseminate (Diesner, Aleyasen, Kim, Mishra, & Soltani, 2013).

3. BACKGROUND ON PRIOR WORK AND ALTERNATIVE SOLUTIONS

Prior work on assessing the impact of documentaries is limited in scope, depth and practical implementations (Barrett & Leddy, 2008; Figueroa, 2002). Major media institutes have proposed

systematic frameworks, which are mainly of theoretical and/or normative nature (Barrett & Leddy, 2008; Clark & Abrash, 2011; Figueroa, 2002; Knight Foundation, 2011). Some frameworks include network related indicators, but fail to implement and measure them. Scholarly work on this topic is primarily confined to studies of psychological effects of films on individuals, and conceptualizes documentaries as a subcategory of mass media.

Overall, evaluation in this domain has typically been done by using (a) traditional, scalable and quantitative methods and metrics, such as the number of visitors of a screening or webpage, and/or (b) conventional, qualitative and small-scale methods for in-depth analysis of the perception of a topic or product by small numbers of people, such as interviews with focus groups. We integrate these two levels by jointly considering (a) the social network of stakeholders involved with the main topic of a movie - whether they have anything to do with a particular production or not - and (b) the substance of the information produced and shared by these groups. The resulting socio-semantic networks allow for reasoning about two types of behavioral information - relationships and information (Diesner, 2013; Roth & Cointet, 2010).

4. BACKGROUND AND METHODOLOGY

Our solution is based on a theoretical framework that we developed by synthesizing indicators of impact based on empirically tested theories from media effects, diffusion research, social and semantic network analysis, and collective action. The resulting CoMTI (content, medium, target, and impact) framework incorporates indicators specific to documentary evaluation that we identified in discussions with subject matter experts as well as additional impact metrics that we considered relevant (Diesner, Pak, Kim, Soltani, & Aleyasen, 2014). This framework considers a variety of stimuli that have been associated with cognitive, attitudinal, and behavioral change on the individual, communal and societal over time. In a nutshell, our methodology involves the following three steps (for details see(Diesner & Rezapour, 2015) (Diesner et al., 2013) :

Baseline model: First we map the current discourse about the main issue addressed in a movie prior to release. This is mainly to understand the existing ecosystem and where impact is possible. Main issues can be identified in a data driven way, e.g. by conducting text summarization techniques on the film transcript, or by filmmakers or funders. Once the main issues are identified, we use a) network analysis to detect key agents and organizations and b) text mining techniques to find trends in current discussions (topics, sentiments, dynamics). For this step, we use ConText (http://context.lis.illinois.edu/) and NodeXL (http://nodexl.codeplex.com/) to collect, analyze and combine text data and network data based on news coverage data, social media data, and interviews with focus groups. Practitioners can use this procedure to understand the given opportunity space for connecting campaign work to relevant stakeholders and themes; helping them to strategically allocate scarce resource and mobilizing social capital.

Ground truth model: This represents the information contained in the actual documentary or media product, i.e. the message that the film can communicate. We understand that there is much more to a film than the actual content, e.g. the cast, images, sound and other aesthetic elements, which are not yet considered with our methodology. At the same time, we argue that the film's content is the smallest common denominator that anybody watching/ listening to a media product could take away. For this purpose, we apply the same text mining techniques as in step one, but this time to the film transcript.

5. ANALYSIS OF MEDIA COVERAGE OF ISSUE AND FILM

a. Data Collection and Curation

To assess the media coverage of the core issue addressed in "Racing Extinction" and the movie itself, we collected and analyzed newswire data from LexisNexis Academic. To map the debate around the main issue addressed in the movie, we consulted with subject matter experts on the film for suitable keywords. Based on their suggestions, we tested various key word combinations and index term constrains for retrieving articles. We divided the search into two time segments: March 23 2014 to January 23, 2015 (= Press Before) and January 24, 2015 to November 24, 2015 (= Press After). The split point reflects the point in time when "Racing Extinction" was screened at the 2015 Sundance Film Festival. We retrieved a total of 30 articles about the movie itself that were published from January 24, 2015 to November 24, 2015.

b. Semantic Networks from Meta Data

The database we construct from the LexisNexis data entails keywords that were automatically assigned and indexed by LexisNexis. The key words represent the main, high-level individuals, organizations, locations and issues addressed in every article if applicable. A percentage value per keyword (also determined by LexisNexis) indicates the strength of association of a keyword with the article. Using ConText, we construct semantic networks from these keywords based on their co-occurrence per article.

Semantic networks are structured representations of information and knowledge that are assumed to represent the knowledge that some person or group has about a topic, and are typically used for reasoning and inference purposes (Diesner & Carley, 2011; Woods, 1975). At a minimum, semantic networks entail nodes, which are also referred to as concepts, and links or edges between the concepts. In this case, concepts represent key words that summarize or synthesize the information provided in news articles, and links are formed for any concepts that co-occur per article. The cumulative link weight represents the number of articles for which a link was observed based on the disambiguated articles. In a semantic network, the meaning of a concept is the ego-network of that concept, i.e. the nodes (alters) and links that get activated when the focal node (ego) is mentioned. The meaning of the entire network emerges from the collective meaning of the nodes as well as the structural properties or patterns of the entire graph. In ConText, the user decides between which categories links shall be formed. For example, connections among and between agents and/or organizations represent a social network. The user also sets the strength value from which on key words are considered; the higher the value the smaller and more focused the resulting network. The networks images below were generated in Gephi (https://gephi.org/) based on the output files from ConText. From the figure 1, we observe that the green color nodes are the most important topics that public are discussing before the movie got released and they include: wild life, climate change, mammals, science and technology, animals, pollution and environmental impacts. Whereas after the movie got released, from figure 2 we observed that there is significant amount of public discourse on the topics such as biodiversity, marine biology, national parks, fresh water ecosystem, science news, global warming and sustainable development. This clearly shows that there is growing public interest on the conservation of wild life and measures that helps in preventing animal extinction. Figure 3 shows the viewpoint of the public and critics on the movie. In this semantic network, the most important nodes are documentary movies, mammals, photography, awards, environmental impacts, and animals that explains the overall concept of the movie.

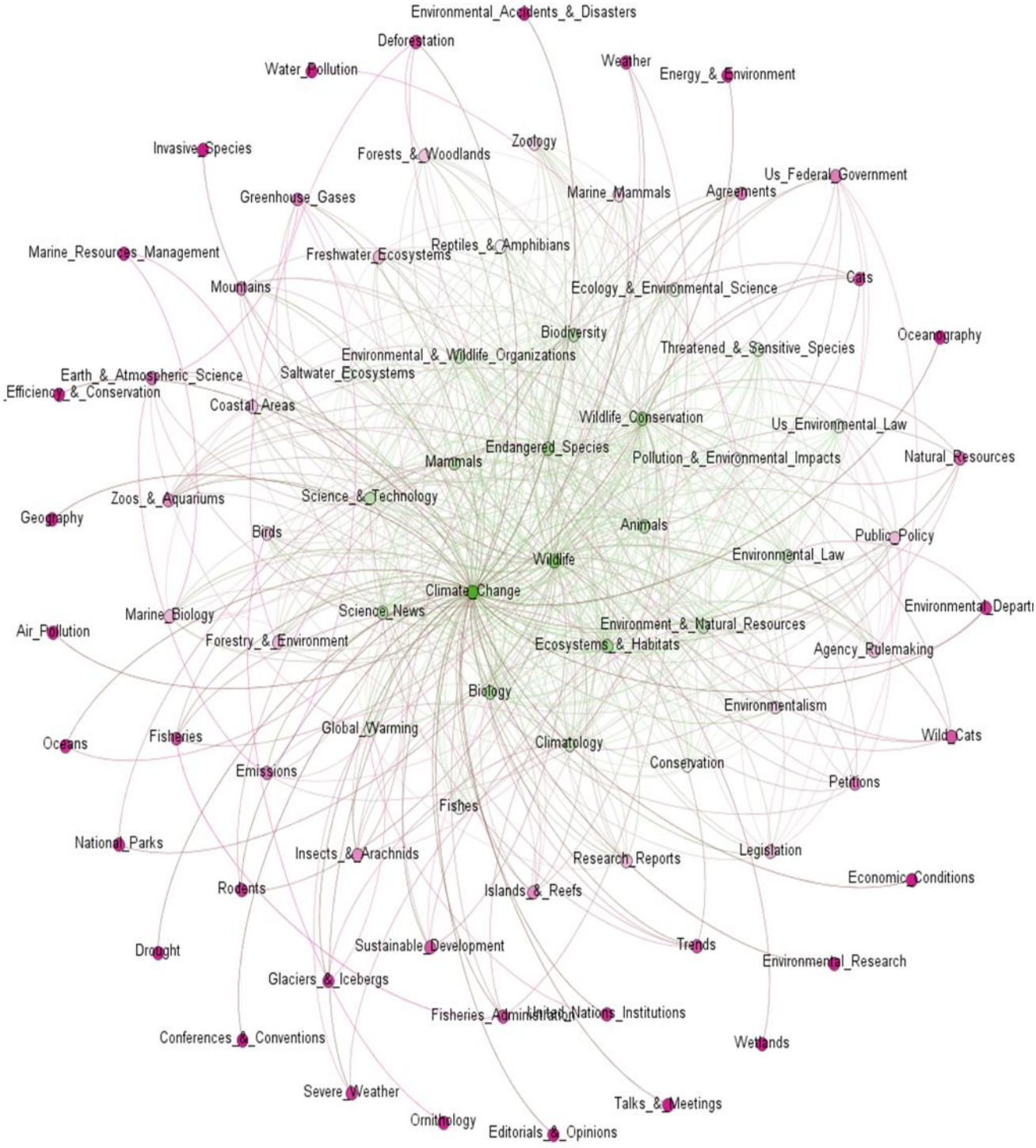

Figure 1: News Coverage before "Racing Extinction" Release: Semantic Network of Meta Data

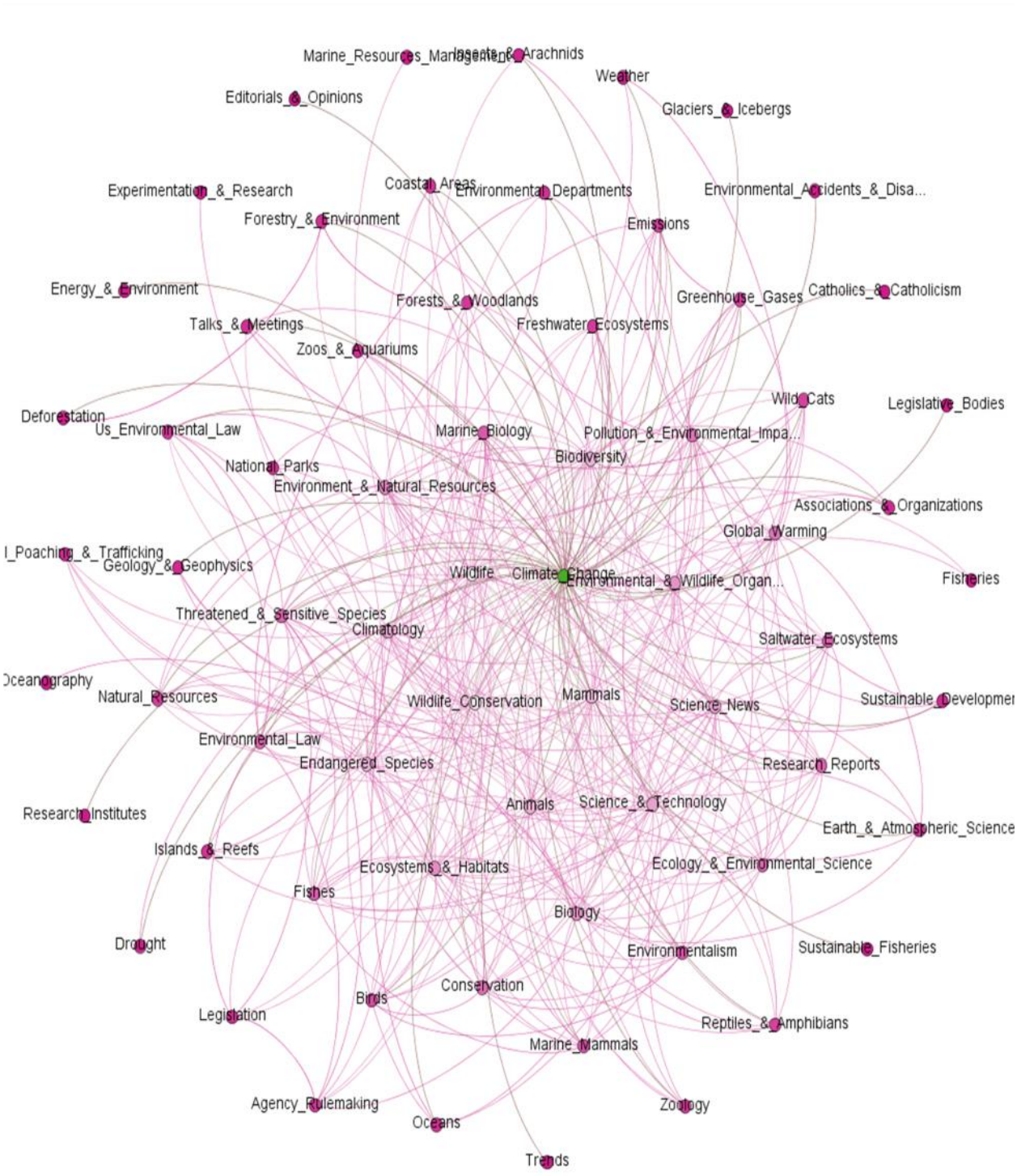

Figure 2: News Coverage after "Racing Extinction" Release: Semantic Network of Meta Data

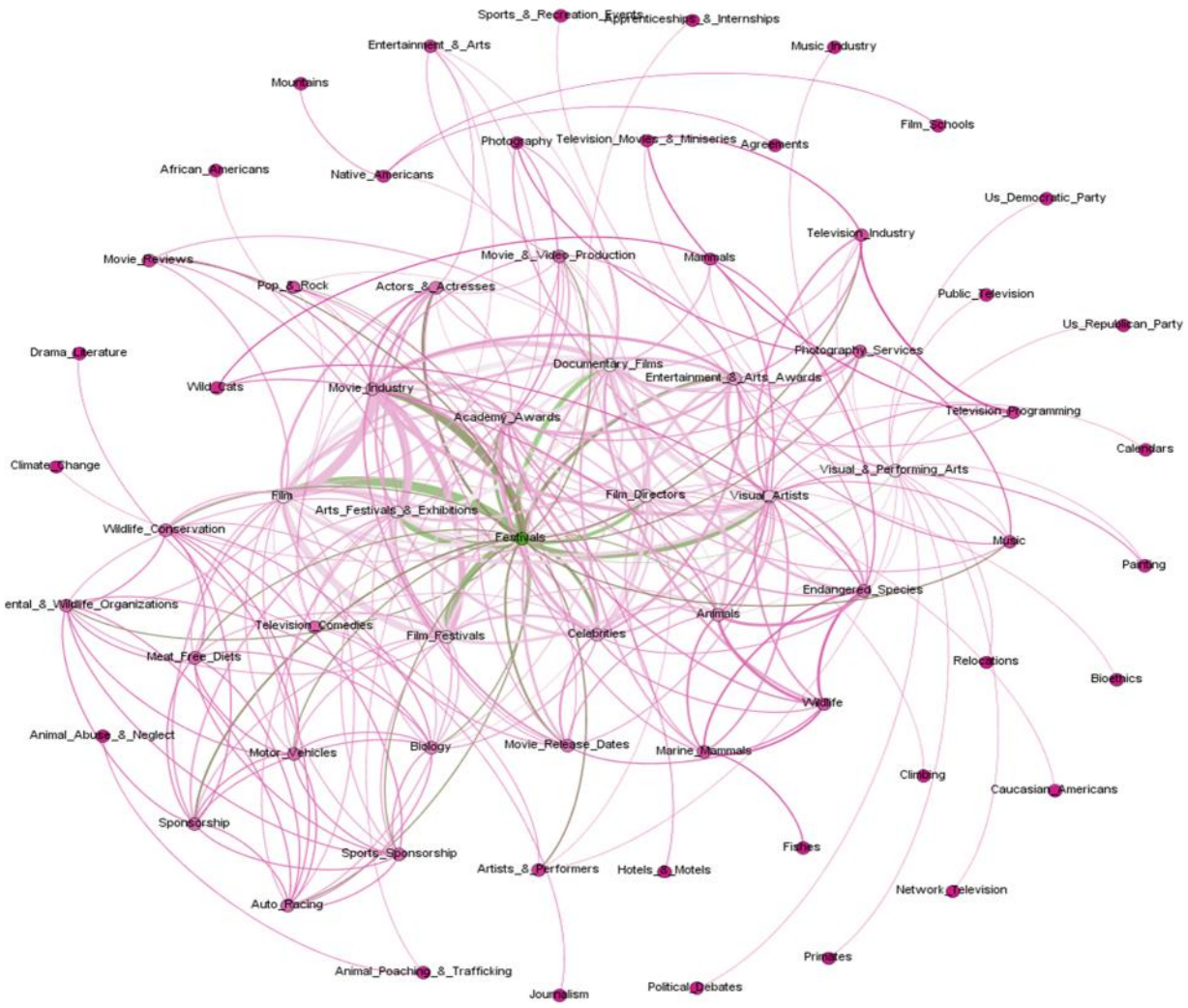

Figure 3: News Coverage "Racing Extinction": Semantic Network of Meta Data

c. Digging Deeper: Summarization of Content of Text Data Sets

After exploring meta-data and keyword networks, we have to dig deeper into the substance of the transcript and media data by analyzing the content of those documents. We use the following three summarization techniques to gain an efficient understanding of the gist of information contained in these different datasets: corpus statistics, topic modeling and sentiment analysis. Basic introduction to each of these techniques is explained below:

Corpus Statistics

One highly effective and efficient way to measure prominence of terms explicitly mentioned in some text data set is to look at cumulative and weighted term frequencies. In short, the terms that are significant and occur in few documents have higher tf-idf scores. By constructing a semantic network by taking tf-idf score, will help us in understanding the text in an efficient manner.

In agreement with the findings from meta-data analysis, the coverage of the issue addressed in "Racing Extinction" is highly focusing on species and habitat information before the release of

the movie. We can see a significant change around the main debate on topics such as climate, population, and conservation in the society after the movie is released.

Term	Frequency	TF*IDF	Ratio of texts occurring in
Species	23,667	9.47E-04	0.85
Habitat	17,390	0.002604	0.54
Information	9,402	0.002861	0.29
Areas	7,406	0.001441	0.45
Critical	7,100	0.002373	0.25
Rule	7,089	0.003417	0.14

Table 1:Salient Terms per Dataset by Weighted Frequency (tf-idf) and Absolute Cumulative Frequency Before the release of the film

Term	Frequency	TF*IDF	Ratio of texts occurring in
Species	13,026	0.001087	0.81
Habitat	6,953	0.00244	0.41
Information	3,624	0.002394	0.19
Climate	3,440	153E-05	0.98
Population	3,208	0.001107	0.42
Conservation	2,695	5.62E-04	0.59

Table 2: Salient Terms per Dataset by Weighted Frequency (tf-idf) and Absolute Cumulative Frequency After the release of the film

After this stage we create a codebook including the most frequent words in the article, and the words with the highest tf-idf score from the meta data list to dig deeper and create a semantic network from the body of the article.

Topic Modeling and Sentiment Analysis

Topic modeling is an unsupervised machine learning technique that summarizes the content of a corpus of unstructured, natural language text data in terms of the most salient topics that are explicitly or implicitly contained in the data (Blei, Ng, & Jordan, 2003; Griffiths, Steyvers, & Tenenbaum, 2007). Each topic is represented by a fit value that indicates how strongly a topic describes a text set, as well as the most salient terms per topic in the underlying data. The terms per topic are sorted by their fit with a topic. By analyzing the appropriate topics in the text, we can better understand the public discourse on “Racing Extinction”.

How do we describe the topic Modeling? Topic modeling is a parameterized method, i.e. the user has to set the number of topics to be identified (we used 7 depending on the size of the dataset), the number of terms per topic (again, 10), the iteration rate for the routine (300), and a list with non-content bearing terms to be excluded from the analysis. We identified the best settings per

data set by running topic modeling with different parameter configurations multiple times and comparing the results.

In the resulting images (Figure 4, Figure 5), each of the most salient topics is displayed as a cluster (typically 5 clusters per Image). These clusters are considered as the different topics addressed in a dataset. The size of a cluster correlates with its prevalence within a single dataset (scaled according to the dirichlet parameter for fit value per topic). The same goes applies for individual terms within clusters: the larger a word, the more strongly it represents the topic it belongs to.

We have also added sentiment analysis into these visualizations: sentiment analysis basically helps to answer the following question: How does an individual or a group of people perceive a topic, product, etc.? The underlying assumption with topic modeling is that human language provides windows to people's emotions and cognition. In the images below, neutral words are shown in blue, words with a negative connotation in red, and words with a positive connotation in green. Words that have no sentiment or opinion value are shown in black.

The press coverage of the movie data before (Figure 4) and after (Figure 5) the release of the film show similar results as that of corpus statistic technique. Public are more interested in studying and discussing more about conservation and wildlife preservation techniques after the release of the movie.

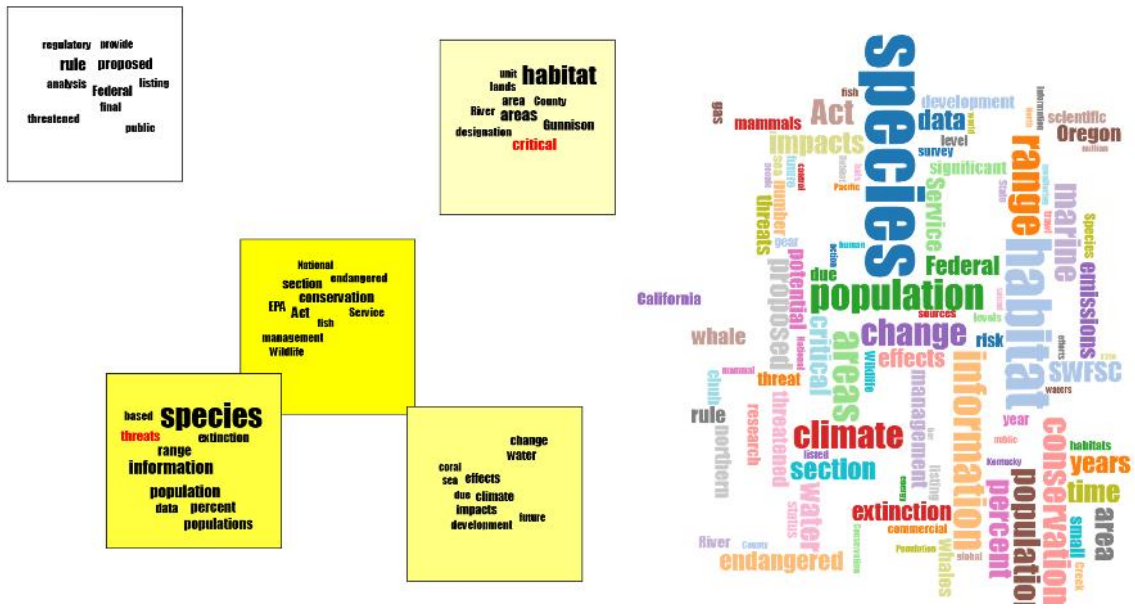

Figure 4: Summary of Press Coverage of Theme before Film Release: Topic Modeling and Sentiment Analysis

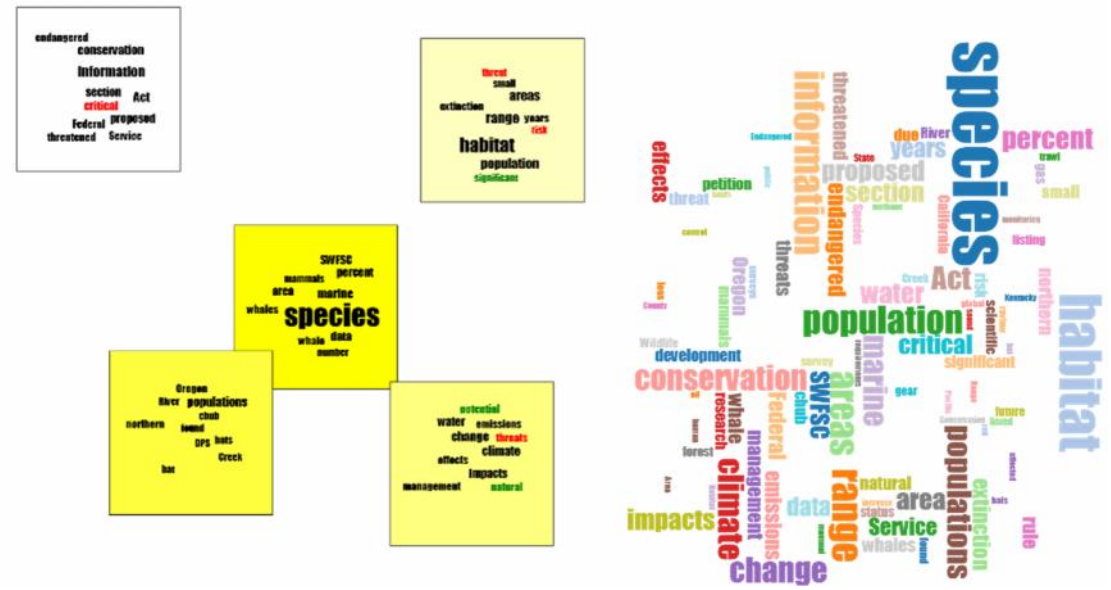

Figure 5: Summary of Press Coverage of Theme after Film Release: Topic Modeling and Sentiment Analysis

6. SOCIAL MEDIA ANALYSIS

We have studied the social media of the movie (Racing Extinction) during November 2015.

a. Twitter:

In order to create the semantic network from the twitter data, we collected the twitter data during the month of November, 2015. From the analysis of twitter semantic data, we obtained the following tables that best explains the semantics of the twitter graph of "Racing Extinction"

Top URLs in Tweet in Entire Graph	Entire Graph Count
https://www.youtube.com/watch?v=-AHbNMXhCF8	30
https://www.thedodo.com/hanako-japan-elephant-61-years-1473985023.html?utm_source=opsFB	29
http://www.discovery.com/dscovrd/racing-extinction-changemakers-leilani-munter/	18
https://docs.google.com/forms/d/161GPc1lQRwj-VN_s6CCmM0OHM_yFrRuEFNXNezXqHPk/viewform	10
http://www.discovery.com/dscovrd/racing-extinction-changemakers-paul-hilton/	9
https://docs.google.com/a/opsociety.org/forms/d/161GPc1lQRwj-VN_s6CCmM0OHM_yFrRuEFNXNezXqHPk/viewform	7

https://actionsprout.io/F2DF8F	6
https://amp.twimg.com/v/08406b56-0c8e-4921-94ec-7d63c5efee3b	5
https://www.youtube.com/watch?v=zdyVOHtpUO4&feature=youtu.be&sf16034991=1	5
http://greenglobaltravel.com/2015/11/25/louie-psihoyos-on-racing-extinction-and-the-cove/?utm_source=hootsuite	3

Top Domains in Tweet in Entire Graph	Entire Graph Count
youtube.com	37
discovery.com	32
thedodo.com	29
google.com	17
twitter.com	9
twimg.com	6
actionsprout.io	6
greenglobaltravel.com	3
racingextinction.com	3
thepetitionsite.com	2

Top Hashtags in Tweet in Entire Graph	Entire Graph Count
racingextinction	123
startwith1thing	70
elephant	30
conservation	3
environment	3
bantrophyhunting	3
opsafarikill	3
sharks	2

somanychoices	1
fuckingshameful	1

Top Words in Tweet in Entire Graph	Entire Graph Count
rt	190
racingxtinction	183
racingextinction	122
startwith1thing	70
gt	52
Years	28
Exhilarating	20
Blue	15
Whale	15

Top Word Pairs in Tweet in Entire Graph	Top Word Pairs in Tweet in Entire Graph
rt,racingxtinction	125
released,exhilarating	31
exhilarating,blue	31
blue,whale	31
whale,feature	31
depressed,elephant	30
elephant,living	30
living,concrete	30
concrete,cell	30

Top Replied-To in Entire Graph	Top Replied-To in Entire Graph
racingxtinction	3
irinagreenvoice	1
vulcaninc	1

Top Mentioned in Entire Graph	Top Mentioned in Entire Graph
racingxtinction	179
gopro	30
shawnheinrichs	23
leilanimunter	21
annekasvenska	18
discovery	17
discoveryuk	7
janegoodallinst	6
seasaver	6
louiepsihoyos	5

Table 3:Twitter Semantic Text Analysis

The above tables explains the top URLs that are mentioned in the twitter graph. The URL for youtube has highest count that explains how the pollution and the climate change are causing the extinction of animal species such as “Blue Whales”. The next highest URL is thedodo, and next comes discovery channel. This shows that people are more interested to share the website link while posting the contents related to “Racing Extinction”. The top hash tags such as racing extinction, conservation, environment shows the concern of public on the extinct species due to environmental pollution. Further, the top words and top word pairs shows the public course on the extinction of animals after the release of the movie “Racing Extinction”.

Semantic twitter graph for Racing Extinction

By considering the top words that many of the users used to tweet about “Racing Extinction”, we tried to develop semantic network that helps us in analyzing the impact of the movie on the public. From Figure 4, we can see that public are showing their concern by using “Emotional” words such as depressed towards extinction and love towards wild life conservation from harmful global warming. Further, cluster network is created to analyze the semantic information in further advanced manner from the twitter data. From figure 5, we understood that there are 7 clusters that explains the twitter data of “Racial Extinction” in an understandable format. First cluster explains racing extinction of different animals, second cluster explains Anneka Svenska, a conservationist in the United Kingdom. Third cluster explains about extinction of blue whales and fourth cluster explains about extinction of wolves in United States of America. Fifth cluster explains about extinction of different species of animals and sixth cluster explains emotions of public towards extinction and environmental impact on the extinction of animal species. Whereas seventh cluster explains about activism towards prevention of animal extinction.

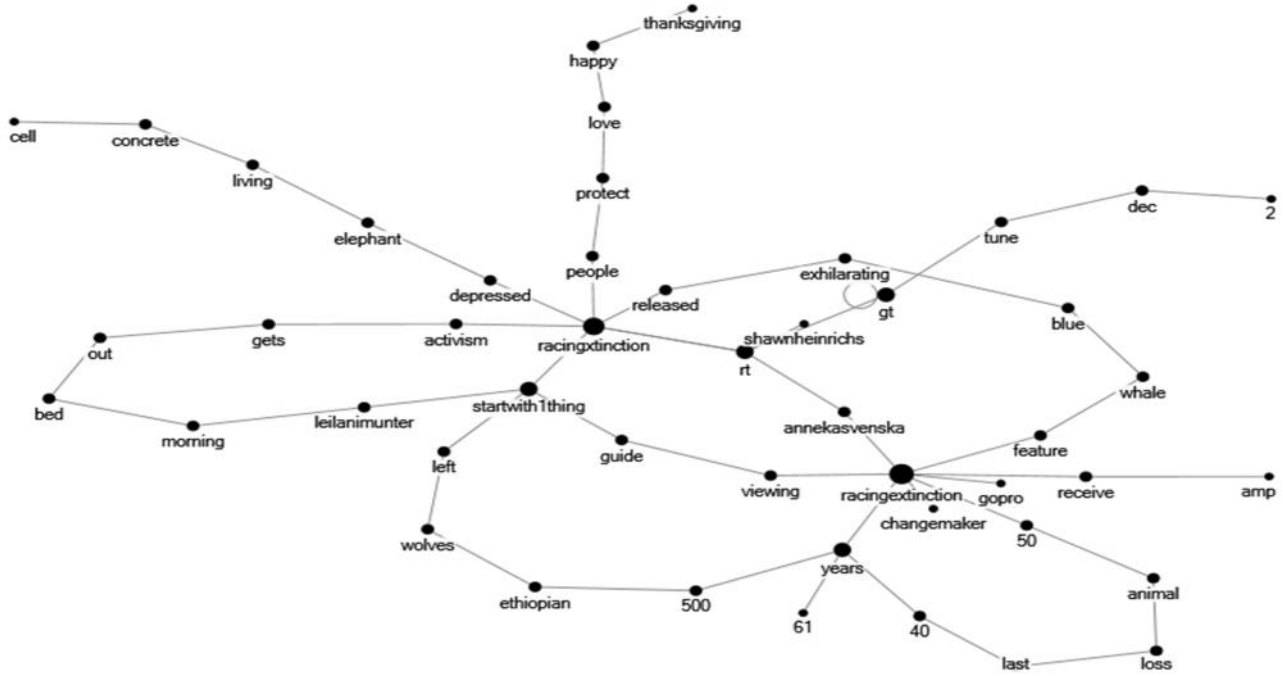

Figure 6: Twitter semantic network for Racing Extinction (November, 2015)

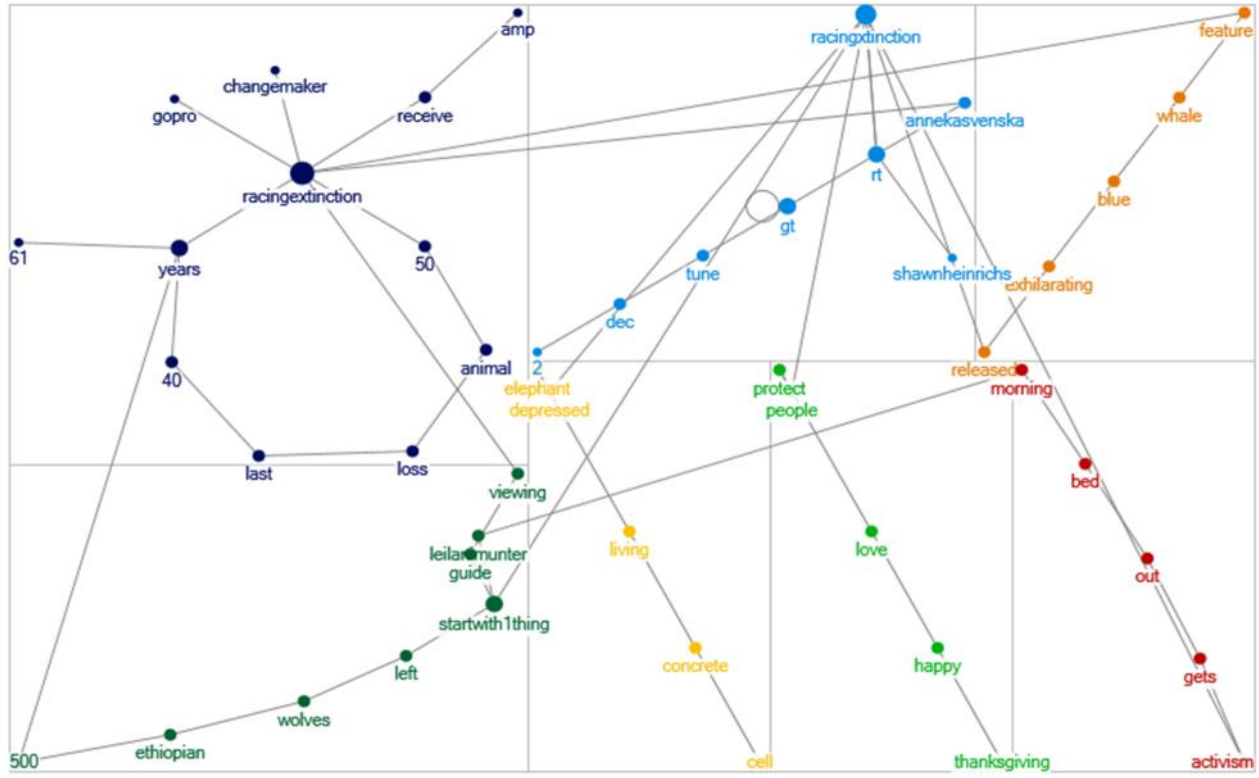

Figure 7: Twitter semantic cluster network for Racing Extinction (November, 2015)

b. Facebook Fan Page:

We do not assume all social media platforms to lead to the same impact; thus, we complement the Twitter analysis with an analysis of user activity on Facebook. On Facebook Fan Pages, users can provide comments to posts. This provides a valuable source for analyzing stimulus (posts) and responses (comments to posts); which together form a public discourse.

"Racing Extinction" movie was telecasted on television in the United States of America on December 2nd, 2015. We collected Facebook data with the time line of one month before the movie telecasted (November 2nd, 2015 to December 1st, 2015) on the television and one month after the movie telecasted on the television (December 2nd, 2015 to January 1st, 2016). We analyzed the data to see the impact of the movie on the public after it got telecasted on the television.

Figure 8 represents the semantic network of the comments from the Facebook fan page data before the movie got telecasted on the television. We had observed that the top nodes are excited, hosting, racing, extinction, and wait. Apart from these nodes, we have nodes such as discovery, channel as top nodes with highest degree, which exemplifies that people are more interested on the telecast of the movie on the discovery channel on December 2, 2015. There are some words such as super, can't, wait that shows the excitement of the public on the telecast of the movie on television.

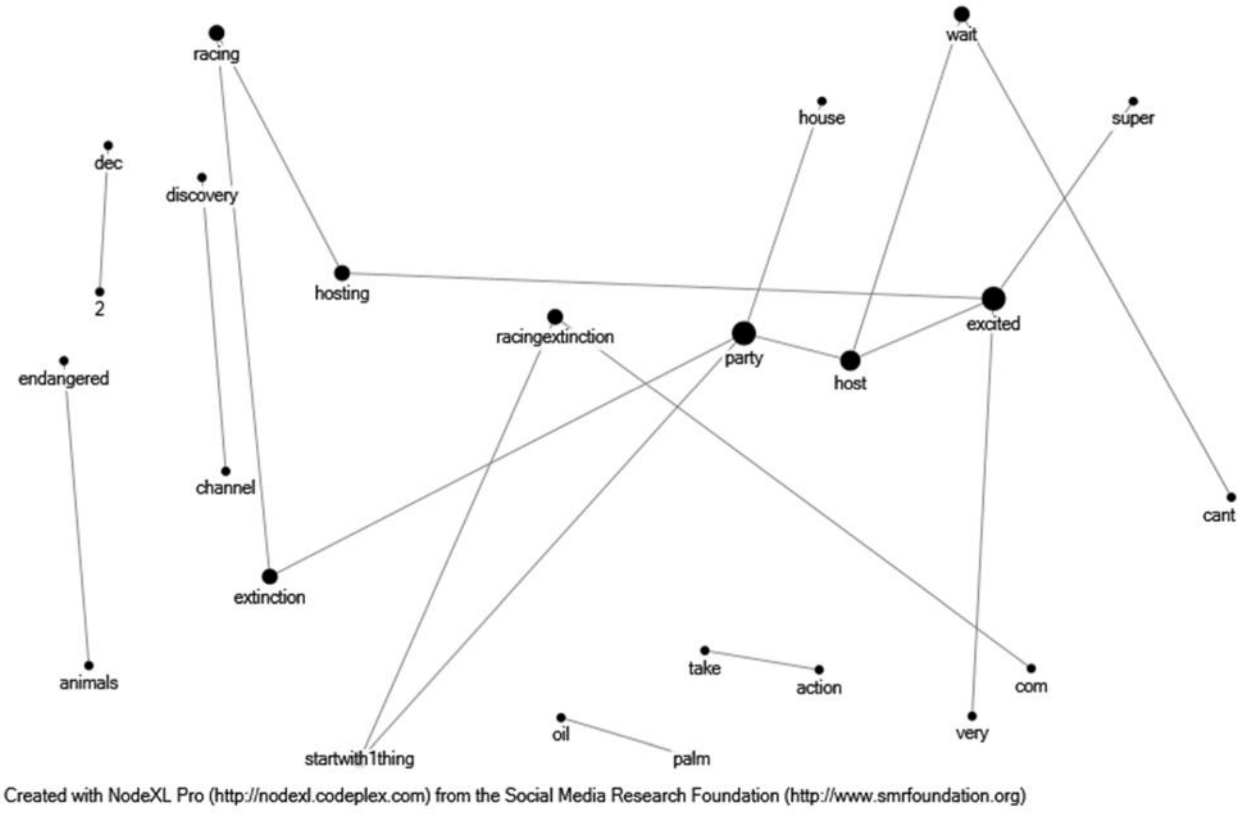

Figure 8: Facebook semantic network for Racing Extinction before television telecast

Figure 9 represents the semantic network of the comments from the Facebook fan page data before the movie got telecasted on the television. We had observed that the top nodes are thing, earth, people, and save. Apart from these key nodes, there are other important nodes such as extinction, endangered, racing extinction, opening, eye, and action which shows the positive impact of the movie on the public after television telecast. Many people are mentioning this movie as an eye-opener and had stated that we should start doing necessary steps in order to save the species that are getting endangered. There are some nodes such as thanks, much which shows that people are thanking for telecasting the movie on the television.
The results has shown that people are more leaning towards discussing about the conservation of wild life and steps to be taken to eradicate global warming. This shows the growing public discourse towards the movie after the movie got telecasted on the Discovery Channel on December 2nd, 2015.

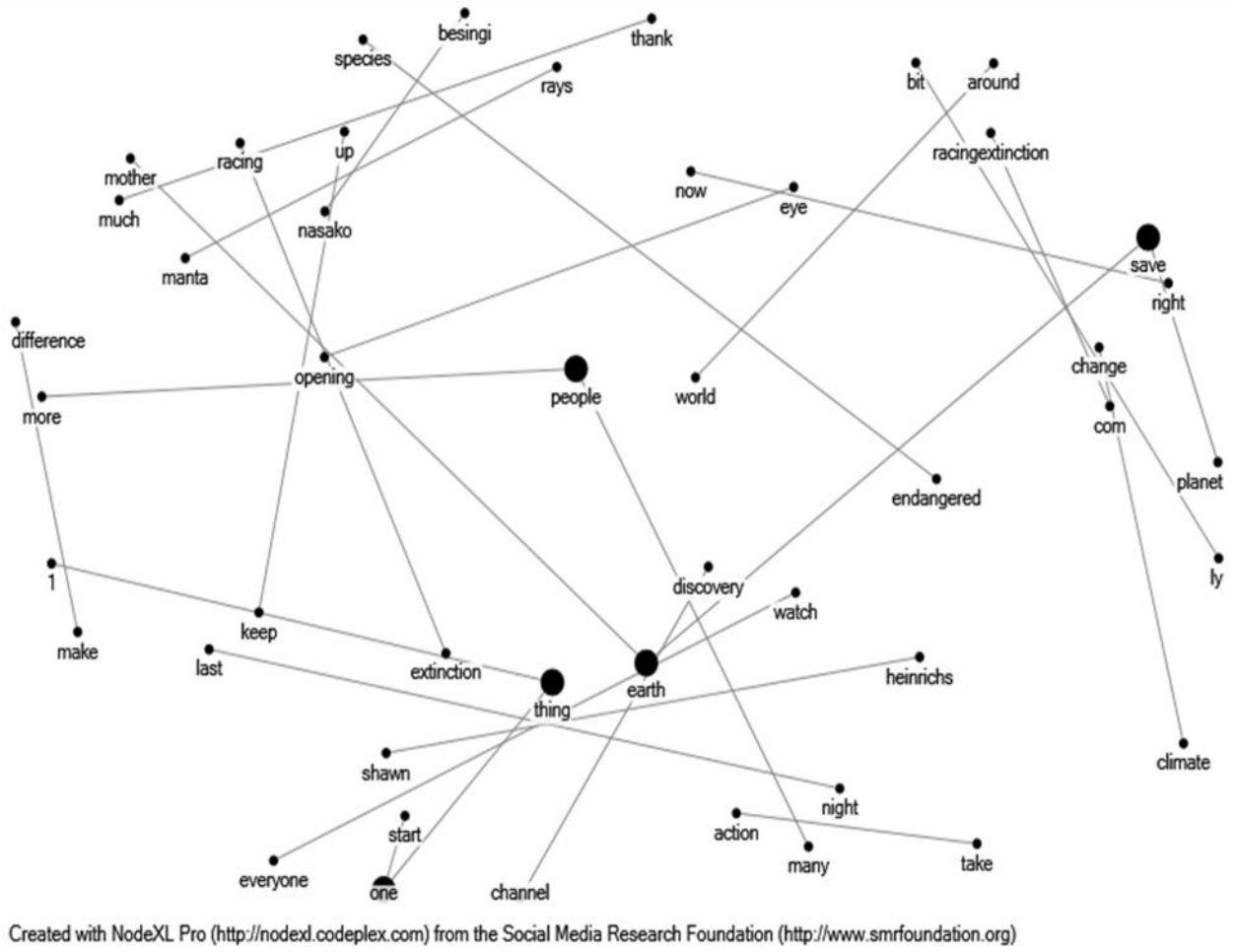

Figure 9: Facebook semantic network for Racing Extinction after television telecast

7. Acknowledgment

This work is supported by the FORD Foundation's JustFilms Fund and Vulcan Production.

8. REFERENCES

Barrett, D., & Leddy, S. (2008). Assessing Creative Media's Social Impact. *The Fledgling Fund*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Clark, J., & Abrash, B. (2011). Social justice documentary: Designing for impact: Center for Social Media.

Diesner, J. (2013). From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data. *Künstliche Intelligenz/ Artificial Intelligence, 27*(1), 75-78. doi: 10.1007/s13218-012-0225-0

Diesner, J., Aleyasen, A., Kim, J., Mishra, S., & Soltani, K. (2013). *Using Socio-Semantic Network Analysis for Assessing the Impact of Documentaries*. Paper presented at the WIN (Workshop on Information in Networks), New York, NY.

Diesner, J., & Carley, K. M. (2011). Semantic Networks. In G. Barnett & J. G. Golson (Eds.), *Encyclopedia of Social Networking* (pp. 766-769): Sage.

Diesner, J., Pak, S., Kim, J., Soltani, K., & Aleyasen, A. (2014). *Computational Assessment of the Impact of Social Justice Documentaries* Paper presented at the iConference, Berlin, Gemany.

Figueroa, M. E. (2002). *Communication for social change: An integrated model for measuring the process and its outcomes*: Rockefeller Foundation.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211-244.

KnightFoundation. (2011). Impact: A Guide to Evaluating Community Information Projects.

Rose, F. (2012). *The Art of Immersion: How the digital generation is remaking Hollywood, Madison Avenue, and the way we tell stories*: WW Norton.

Roth, C., & Cointet, J. (2010). Social and semantic coevolution in knowledge networks. *Social Networks, 32*(1), 16-29.

Woods, W. (1975). What's in a link: Foundations for semantic networks. In D. Bobrow & A. Collins (Eds.), *Representation and Understanding: Studies in Cognitive Science* (pp. 35-82). New York, NY: Academic Press.

Diesner, J., & Rezapour, R. (2015). Social Computing for Impact Assessment of Social Change Projects *Social Computing, Behaviosral-Cultural Modeling, and Prediction* (pp. 34-43): Springer.