Assessing topic discovery evaluation measures on Facebook publications of political activists in Brazil

Arian Pasquali
LIAAD - INESC TEC, DCC - FCUP, Universidade do Porto, Portugal
arrp@inesctec.pt

Marcela Canavarro
FEUP, Universidade do Porto - PEICC/UFRJ, Brazil
mcanavarro@gmail.com

Ricardo Campos
LIAAD - INESC TEC, I.P. Tomar, Portugal
ricardo.campos@ipt.pt

Alípio M. Jorge
LIAAD - INESC TEC, DCC - FCUP, Universidade do Porto, Portugal
amjorge@fc.up.pt

ABSTRACT

Automatic topic detection in document collections is an important tool for various tasks. In particular, it is valuable for studying and understanding socio-political phenomena. A currently relevant example is the automatic analysis of streams of posts issued by different activist groups in the current Brazilian turmoil, through the analysis of the generated streams of texts published on the web. It is useful to determine the relative importance of the different topics identified. We can find in the literature proposals for measuring topic relevance. In this paper, we adopt two of such measures and apply them to data sets extracted from Facebook pages related to Brazilian political activism. On top of the analysis, we then carry an experimental evaluation of the human interpretability for these two measures by comparing their outcomes with the opinion of three Brazilian professionals from the field of Communication Science and media-activists.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis; J.4 [**Social and Behavioral Sciences**]: Sociology

Keywords

Computational Social Science, Natural Language Processing, Web mining, Topic Modeling, Coherence Evaluation, Computational Linguistics

1. INTRODUCTION

Traditionally, there has been substantial interest within computational linguistics in techniques for exploring a large volume of text documents without autonomous analysis. Social networks like Twitter and Facebook have become usual places where people share billions of posts every day. Among the stream of status updates, lies valuable information about their opinion in a variety of subjects. Politics is one of them.

The interest in mining the Web data for political insights has increased since the booming of popular upheavals around the world, in the 2000's, especially after the Arab Spring. Many authors [9],[12],[17] and [4] agree that recent uprisings have been a result of a complex network of interactions both on social networks and live political demonstrations (sometimes simultaneously). On these grounds, some researchers have begun to explore open social data to study topics on social science [30] [14] [16] [26].

Probabilistic topic modeling [6] is a popular approach to explore textual document collections with no prior annotation. Topic modeling can provide a summary of the document collection that would be impossible to obtain by hand and may reveal connections between and within documents that were not evident. Due to these traits, they are frequently used as an exploratory tool [11] [20] [27].

As pointed out by D. Blei, despite the popularity of topic modeling, the unsupervised nature makes it hard to validate its results [6]. Topics learned by the model are subject to subjective judgment. Two experts can easily disagree about the meaning and/or usefulness of a topic [24].

In this work, we apply topic modeling on Facebook posts in Portuguese related to political movements and verify whether two automatic evaluation measures can model human judgment when working with short and badly structured texts. In particular, our aim is to assess the compliance of the measures with external human evaluation. We begin with a brief overview of the simplest probabilistic topic model algorithm, Latent Dirichlet Allocation (LDA) [7]. Then we describe in detail two measures that aim to evaluate the coherence and cohesion of learned topics (Extrinsic UCI and Intrinsic UMass) automatically. In the following section, we describe how we applied these two measures to a real-world data set collected from Facebook. We then discuss the outcomes of the experiment and present our conclusions. The paper also illustrates how text/web mining techniques can be used to explore new paths in the treatment of a large

C3S2E '16, July 20-22, 2016, Porto, Portugal

http://dx.doi.org/10.1145/2948992.2949015.

amount of text in the field of Humanities, such as Political Science and Communication.

2. TOPIC MODELING

The main goal of topic modeling is to identify the topics from a collection of documents automatically. Latent Dirichlet Allocation [7] (LDA) is a well-known topic modeling algorithm. It is based on the idea that a document was generated by picking a distribution of topics, and words.

Documents can be represented as a mix of topics, where a *topic* is defined as a 'distribution probability over words in a vocabulary'. The resulting topics can be presented to the user as a list of words that appear with high probability in that topic. As described by Toshiaki et al. (2014) [15], LDA generative process for a document can be described as follows:

For each word in the document,

- Choose a topic t according to the probability distribution over topics that the document has;
- Choose a word w according to the probability distribution over words that topic t has.

Usually, the simplest version of LDA holds no prior information about the documents, which means that they are not labeled with topics or keywords. Evaluating topic models is not trivial, and it is an active field of research [28]. The challenges of evaluating this kind of method are discussed in detail in the next session.

3. RELATED WORK

3.1 Computational Social Science

Dealing with a massive amount of data is not new for political studies. Nevertheless, the advent of social media and its recurring influence on popular upheavals in the 2000's brought up new challenges for social scientists. It is not clear how far this influence goes, but a number of authors have been focusing on social media data sets to understand recent events such as the Arab Spring (2010), the *Indignados* movement (Spain, 2011) and the *Journeys of June* (Brazil, 2013). Provide them a set of effective tools for mining all this data should be a relevant task to technologists and computing engineers.

Even before the explosion of the Arab Spring and following popular upheavals around the world, the social scientist Bruno Latour recognized 'digital traces' as an open door for social researchers to access 'inner workings of private worlds' through 'their inputs and outputs [that] have become thoroughly traceable' [18].

Manuel Castells [9] analyzed the 2000's popular riots based on the idea that people's outrage is now linked through digitized networks, enabling practices of self-communication[8] instead of relying mainly on the traditional mass-media. Inspired by Castells, the *Network, Movement and Technopolitics* research group led by Javier Toret has gone deep into Twitter datasets in order to understand and explain the *Indignados* movement in Spain. By analyzing the spread of ideas within the studied networks Toret[12] and his team have proposed that the Spanish upheaval has followed the Arab Spring - and has also grown itself - through a process of digitally-enabled *emotional contagious*, which can be detected on networks visualizations by data mining Twitter data sets.

Ben-David and Matamoros-Fernández [3] applied topic learning to analyze a dataset extracted from Facebook pages in order to study hate speech on social media while Warner and Hirschberg[29] relied on comments extracted from Yahoo! groups and the American Jewish Congress to apply data mining techniques to study the same thematic.

In Brazil, the Laboratory for Studies on Image and Cyberculture (Labic) has been intensively studying the 2013 Brazilian mobilization known as "Journeys of June" through Twitter datasets. Labic has also been following the 2013 event's influence on the country's current political situation.

3.2 Topic Coherence Evaluation

Until recently, evaluation of topic models had focused on statistical measures of perplexity or likelihood of test data [28]. According to Wallach et al. (2009), topic models are better evaluated by either measuring performance on some secondary task, such as document classification or information retrieval or by estimating the probability of unseen documents given some training documents (applying the Perplexity measure) [28]. In this context, we can use labeled data to evaluate model performance in terms of precision and recall [13], but this is not always practical and demands prior annotation by an expert on the domain of the documents.

However, as described by L. Alsumait et al. (2009), not all learned topics are of equal importance or correspond to genuine themes of the domain [2]. Some topics can be a collection of irrelevant words or represent insignificant topics, meaning that a topic can hold no semantic value for the user and still be statistically valid. In fact, as demonstrated by Chang et al. (2009) [10], conventional topic evaluation measures, like those proposed by Wallach et al. (2009), do not necessarily reflect the semantic coherence of individual topics learned by a topic model, making it difficult to evaluate how well a topic model will perform in some end-user task. In Chang et al. (2009) experiments, it was proved that sometimes perplexity could be contradictory to human judgment regarding interpretability of the learned topics.

Many authors in the Natural Language Processing (NLP) community have proposed coherence measures to evaluate topics learned by a given topic model. While topics learned by topic models often look useful, sometimes that is not the case [22]. Automatically quantifying topic coherence helps to identify 'junk' topics that may be statistically well founded, but meaningless to end users. Automatically identify 'junk' topics can lead to better ways to interact and explore the data [22]. For instance, David Newman et al. (2010) published other study [23] about automatic evaluation of topic coherence. Their main contribution was to present a measure based on Point-wise Mutual Information (PMI) to model human scoring which has been shown to be highly correlated with human evaluation. Their method relies on co-occurrence of words in an external reference source such as Wikipedia for automatic evaluation of topic interpretability.

Later, David Mimno et al (2011) proposed similar method without using external source [19]. As described by [1], their method defines topic coherence as the sum of the log ratio between co-document frequency and the document fre-

quency for the N most probable words in a topic. The intuition behind this method is that the co-occurrence of words within documents in the corpus can indicate semantic relatedness.

In this work, we implemented and tested the intrinsic measure UMass [19] proposed by Mimno et al. (2011), and the extrinsic measure UCI [21] proposed by Newman et al. (2010).

4. FORMAL DEFINITION

A topic coherence measure scores a single topic by measuring the degree of semantic similarity between high-scoring words in the topics. These measures help in distinguishing topics that are semantically interpretable from topics that are a result of statistical inference. Topic coherence is defined as *sum* of a particular coherence score for each pair of representative words. The following Equation 1 is the general definition of coherence and was inspired by [24].

$$\text{TopicCoherence}(T) = \sum_{w_i \in V; w_j \in V} score(w_i, w_j, \epsilon). \quad (1)$$

Where:

- T is topic (i.e. a set of words describing T);
- *score* is a function responsible for measuring the coherence between a pair of words;
- w_i and w_j represents a pair of words;
- V represents the whole vocabulary;
- the term *epsilon* can be used as smoothing value depending on the nature of the dataset and prevents the occurrence of extreme values.

4.1 Intrinsic UMass Measure

The UMass measure [19] is based in the co-ocurrence of words within a given document, computing the correlation of words in a given sliding context window. The measure is defined as follows:

$$score_{Umass}(w_i, w_j) = log \frac{D(w_i, w_j) + 1}{D(w_i)}, \quad (2)$$

where *D(w_i, w_j)* counts the number of documents that contain words w_i and w_j and *D(w_i)* counts the number of documents containing w_i. The UMass metric is computed within the original data set rather than an external corpus source like UCI. It is an intrinsic metric by nature and aims to confirm that the topics and words selected by the model are known to be in the data set.

4.2 Extrinsic UCI Measure

The UCI measure [21] is based on a Pointwise Mutual Information (PMI) pairwise score function:

$$score_{UCI}(w_i, w_j, \epsilon) = log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)}, \quad (3)$$

where *p(w)* represents the probability that w_i is present at a random document and *p(wi,wj)* represents the probability of both w_i and w_j being present in the same document. This probability is estimated by using an external knowledge data set such as Wikipedia [21]. The term *epsilon* can be used as smoothing value depending on the nature of the dataset and prevents the occurrence of extreme values. The probabilities are calculated as follows:

$$p(w_i) = \frac{D_{Wikipedia}(w_i)}{D_{Wikipedia}} \quad (4)$$

and

$$p(w_i, w_j) = \frac{D_{Wikipedia}(w_i, w_j)}{D_{Wikipedia}}; \quad (5)$$

where $D_{Wikipedia}$ counts the number of documents at Wikipedia containing the word. UCI can be regarded as an external source to compare the words present in a given document with an already existent set of topics/words that gathers accumulated subjective semantic evaluations.

5. THE EXPERIMENT

We applied topic modeling on political messages published on 36 Facebook pages and then asked three annotators to analyze the relevance of each learned topic. Their scores were compared to the UMass and UCI scores in order to compare the automatic evaluation with the human judgment. We then analyzed the outcomes considering their utility to the field of text mining and their application in the area of political studies.

The 36 considered pages are a fraction of a 320-page network under investigation as part of the Ph.D. of one of this study's authors. Therefore, the pages were selected and categorized into six different classes based on previous knowledge about their general features/profile, and according to the following criteria:

- Show some relevance in the production and/or dissemination of content on Brazilian contentious politics;
- Not being a social network official page of any corporate media organization;
- Be identified as a collective identity instead of a singular individual (a recurring feature in political actors in social media);
- Be active between March 1, 2015, and February 29, 2016 (data collection time range).

5.1 The Data Set

All data was collected in mid-March, 2016, using the application Netvizz 1.25 [25] which retrieved 314,973 documents (posts, photos, videos, link shares and events) for the 36 pages together. Each page generated one tab file, and the six files of each class were considered together, as a unique data set, which means that the experiment runs into 6 data sets - one per each 6-page class.

Netvizz lets the researcher choose between the last N posts and all posts published in a window of time. We opted to collect data from March 1, 2015, to February 29, 2016, because that was an intensive period in the Brazilian political context, generating lots of relevant content in social networks. Then we run the application to retrieve the data automatically.

The generated data set aggregates 313,514 posts, considering each status update, photo, video, note and link share on a page as a *document* (note that 1,459 *Facebook events* were ignored in the total of documents as they were not in the scope of this study). Each class' features are described bellow (Table 1 lists all pages considered and Table 5.1 refers their general features).

Class 1 - social movement with a singular main cause: page focused on a specific kind of Right, disseminating topics related to its main cause. It is managed by activists who maintain actions on the streets and on digital social networks.

Class 2 - grassroots media: leftist collective identities[1] that disseminate own-produced and third-party news pieces, mainly about social movements, popular demonstrations and other related topics. Many of them were born from massive popular protests in Brazil in 2013 and tend to be neither pro-President Dilma Rousseff nor pro-impeachment. They are frequently confronting mass-media outlets' versions on political topics.

Class 3 - Pro-President Dilma Rousseff administration: news outlets that disseminate own-produced pieces. They also tend to share lots of content from each other and are frequently confronting mass-media outlets' versions on political topics.

Class 4 - Rightist news outlets that disseminate own-produced and third-party pieces that demand president Dilma Rousseff impeachment. They are also consistently against left-wing administrations in other Latin American countries and adopt a strong discourse against corruption.

Class 5 - Rightist pages that spread viral *memes* and third-party links demanding President Dilma Rousseff impeachment. They are frequently against left-wing administrations in other Latin American countries and are more focused on easy-to-turn-viral content than analytical or descriptive news pieces.

Class 6 - Pages with a progressivist view of political themes. They are more focused on easy-to-turn-viral content than analytical or descriptive news pieces although sometimes they publish third-party news, usually with sarcastic comments.

It is important to note that none of the pages is a Facebook page of any traditional media corporation and all of them are named as a collective identity instead of a singular individual's profile (either a public person or a common user). Nevertheless, a few of them might be managed and updated by only one person.

Table 1: Facebook pages

Class	Facebook Page Name
Class 1	- Aliados do Parque Augusta - Comitê Popular Rio Copa e Olimpiadas - Das Lutas - Garis do Rio de Janeiro em Luta - Movimento Passe Livre - Ocupe Estelita
Class 2	- A Nova Democracia - Guerrilha GRR - Mariachi - Midia Independente Coletiva - MIC - Papo Reto - Vírus -Planetário
Class 3	- Brasil 247 - Diario do Centro do Mundo - Favela 247 - Revista Forum - Jornal GGN - Pragmatismo Político
Class 4	- Correio do Poder - Folha Política - Implicante - O Antagonista - O Reacionário - Vem Pra Rua Brasil
Class 5	- Humor 13 - Movimento Brasil Livre - Movimento Contra a Corrupção - Movimento Endireita Brasil - TV Revolta - Revoltados Online
Class 6	- Acorda Meu Povo - Deboas na Revolução - Movimento Pro-Corrupção - O Badernista - Porque Eu Quis - Rede Esgoto de Televisão

5.2 Methodology

The methodology applied in this study combines the application of LDA using collapsed Gibbs sampling, the computation of Intrinsic UMass and Extrinsic UCI scores, and human evaluation. It aims to test if automatic evaluation from learned topics proves to be useful in identifying coherent topics in datasets with short and badly structured text in Portuguese. Inspired by D. Newman (2010) methodology, we defined an experiment to evaluate the correlation between human judgment regarding observed cohesion against the intrinsic measure UMass and observed coherence against the extrinsic measure UCI. We understand that there is an important difference on what intrinsic and extrinsic measures try to model. Cohesion measures how much the words representing a specific topic have in common without any source beyond the original document collection used to build the model, and it corresponds to UMass assumptions. Coherence, on the other hand, quantifies if there is any semantic meaning between the words that represent a topic, which corresponds to UCI assumptions. We implemented these measures and calculated their correlation with human judgment, as described bellow:

a) Selection of important data within the Facebook posts data set.

We kept only the original text of each publication and the type of post (status update, link and photo). All other data were not considered (e.g.: users unique number ID; number of likes, comments and shares; post ID; date of publication).

b) Build local Wikipedia index with entries in Portuguese.

We built the index using a dump provided in March 2016

[1]For detailed information on the concept of "collective identities", see TORET[12].

Table 2: Number of posts per class of page

Class	*Features*	*Posts*
1	Particular cause (Social Movement)	7,367
2	Grassroots news (Leftist)	14,591
3	Pro-governism news (Center)	47,080
4	Pro-impeachment news (Rightist)	37,433
5	Pro-impeachment virals (Rightist)	196,641
6	Progressivist virals	10,333
TOTAL	-	313,514

by Wikipedia at https://dumps.wikimedia.org [2]. A total of 2,065,963 of documents in Portuguese were considered. This index was used to compute the extrinsic coherence measure.

c) Implementation of standard procedures for text pre-processing.

In order to tokenize and pre-process [3] each document, we have implemented standard procedures [5]. We considered each post of each page as one document. In this phase, the outcome was a matrix that showed the n most frequent words (uni-grams) in each document in the data set. This generated a bag-of-words representation on which we applied LDA.

d) Application of topic modeling algorithm.

This phase consisted in the application of LDA using Gibbs sampling to learn 15 topics from each class. Later, we selected the 9 words that were most likely to appear in a given topic. The number of topics were empirically defined, we use the same number of topics for all 6 classes. In this experiment we were only interested in highlighting good and bad topics in terms of interpretability; finding the optimal number of topics for each class was out of the scope of this work.

e) Computation of UMass and UCI measures.

This phase generated scores to evaluate the interpretability for each of the 15 learned topics from the previous step. It worked as follows:

- *Extrinsic UCI score:* shows an extrinsic measure, computed by comparing the documents' words distribution probabilities with Wikipedia's articles' words distribution probabilities. The pair-wise probability was considered when the words appeared on an interval of 10 on a particular document.

  Possible scores range from real values greater than 0 to infinity, meaning that topics with higher values are evaluated as better than topics with lower values. Since the data set was mainly in Portuguese, we downloaded and built a local index with Wikipedia articles in Portuguese and used it as the external reference to support UCI.

- *Intrinsic UMass score:* shows an intrinsic measure, computing words distribution probabilities among all the words that the documents contain. The pair-wise probability was considered when the words appeared on an interval of 10 on a particular document. Possible scores range from real values greater than 0 to infinity, meaning that topics with higher values are evaluated as better than topics with lower values.

f) Human evaluation.

All three annotators were familiar with the general thematic on the pages (one annotator is one of the authors of this paper). They are professionals in the Communications field and are personally involved in the Brazilian political scenario to which the pages' content relates to. Each annotator has analyzed all the 15 learned topics with 9 words for each class of pages (see Table 3), indicating:

- *Cohesion* score (from 1 to 5) among the 9 top words of each topic, where 1 is the lowest level of cohesion and 5 the highest one. They were asked to analyze whether the words in a given topic showed a sense of unity, that means, if, accordingly to their knowledge and experience on the general thematic, they could see relevant connections among those words.

- *Coherence and comprehension* score (from 1 to 5) within the 9 top words of each topic, where 1 is the lowest level of coherence and comprehension, and 5 the highest one. They were asked to analyze to which extent it was possible to understand the general thematic, through logical interconnections among the words themselves but also between those words and subjacent ideas that the annotators previously knew about the thematic.

g) Comparison of annotators scores with UCI and UMass.

We ranked the 15 learned topics for each class of pages accordingly to their UMass scores (from highest to lowest) and compared it with their respective cohesion score set by the annotators. Our goal was to compare the ranking set by the *intrinsic measure for automated evaluation* with a *human judgment of internal cohesion* of the learned words in each topic.

We then ranked the 15 learned topics for each class of pages accordingly to their UCI score and compared it with their respective coherence/comprehension scores set by the annotators. Our goal was to compare the ranking set by the *extrinsic measure for automated evaluation* with a *human interpretability and interconnection of hidden ideas subjacent to the known words* in the learned topics. The comparison task was done computing Spearman correlation between automated and human annotated scores.

h) Computation of inter-rates agreement.

We computed how much the annotators agree with each other to validate the evaluation. Given the subjectivity of the task we have grouped the two lower levels of rating and the two upper levels for assessing inter-rater agreement. This resulted in a three level scale.

5.3 Outcomes

We present in this section the outcomes in terms of the correlation between human judgment and automated evaluation. We applied Spearman correlation for this task.

[2]Portuguese Wikipedia dump downloaded from https://dumps.wikimedia.org/ptwiki/20160305/ptwiki-20160305-pages-articles-multistream.xml.bz2

[3]Such as filtering out articles, pronouns, prepositions, certain kinds of verbs (e.g.: to be, to have), Facebook update messages, files terminations (.jpg, .png, .gif) and the words yes/no. Part of our filtering/stopword process was based on *Snowball: A language for stemming algorithms* and *Ranks NL Stopwords Portuguese* by Damian Doyle

Table 3: Annotators tasks example

Words	*Cohesion*	*Comprehension*
rio, esquerda, professor, janeiro, paulo, carlos, universidade, partir, centro	2	2
brasil, governo, povo, presidente, federal, direitos, direito, poder, caso	4	5
garis, greve, trabalhadores, luta, comlurb, sindicato, rio, gari, chapa	5	5
transporte, aumento, copa, movimento, mundo, governo, passe, livre, tarifa	4	4
povo, negro, marcha, reaja, campanha, internacional, anos, luta, dia	3	5
ato, dia, policiais, rio, pessoas, protesto, frente, apoio, rua	2	3
bem, pessoas, coisa, cidade, sempre, poder, fazendo, anos, bom	1	1
direitos, rio, dia, humanos, janeiro, ativistas, mil, caso, segundo	2	3
movimento, dia, coletivo, popular, movimentos, rede, social, luta, coletiva	1	2
apoio, moradores, prefeitura, vila, luta, hoje, solidariedade, novas, praia	2	4
povo, anos, pior, pessoas, banco, hoje, dias, brasileiro, infelizmente	1	1
parque, pic, nic, circulo, dia, poder, cidade, gente, podemos	3	4
rio, vila, moradores, prefeitura, projeto, comunidade, prefeito, eduardo, copa	2	5
parque, augusta, cidade, municipal, prefeitura, dia, luta, rua, guarda	4	5
mulheres, pessoas, sociedade, forma, vida, mulher, nunca, grupo, homens	3	2

5.3.1 *Comparison between UCI and annotators coherence/comprehension score*

We can clearly see that the annotators have very similar correlations with the automatic measure for all classes when compared to UCI results, with the exception of class 5 (see Figure 1, where bars represent the three different annotators). In this case, two of the annotators tend to disagree with UCI. In order to assess the agreement of the annotators we have also calculated Fleiss' kappa, a usual measure of inter raters agreement. In this case the values of kappa range between 0.209 and 0.53 for all the classes but 5, where it is negative (Table 4). Being 1 maximum agreement and -1 maximum disagreement, we see that there is a moderate concordance between raters in all classes but one. The low p values indicate that the value of kappa for that class is significantly different from zero.

Table 4: Coherence inter-raters agreement

Class	1	2	3	4	5	6
kappa	0.265	0.43	0.53	0.453	-0.009	0.209
p value	0.015	0.00	0.00	0.000	0.938	0.053

5.3.2 *Comparison between UMass and annotators cohesion score*

Table 5: Cohesion inter-raters agreement

Class	1	2	3	4	5	6
kappa	0.061	0.296	0.317	0.138	0.225	0.231
p value	0.575	0.006	0.003	0.211	0.038	0.032

Concerning UMass, we can observe that the correlation between annotators' rates and UMass always go in the same direction class-wise (Figure 2). This tendency for agreement is confirmed by the positive values of kappa (Table 5). However two of the p values are well above 0.05, a usual level of significance for considering that the value of kappa is significantly above zero. In the case of UMass, classes 1 and 4 show the highest disagreements between annotators and automatic measure.

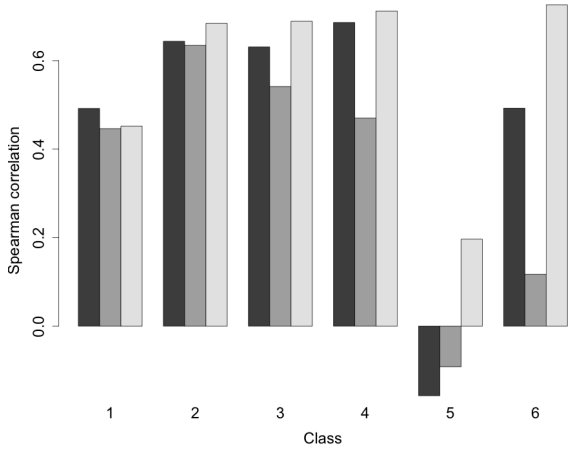

Figure 1: Correlation between UCI and annotators

6. DISCUSSION

It is important to note that the authors who proposed UCI [22] and UMass [19] have tested their results against datasets with well-structured text, such as news and academic papers. In this work we faced a variety of texts from Facebook posts which are usually short and not necessarily well written or structured.

The correlation between UCI and human scores for the Class 5, as shown in figure 1, presents an interesting behavior, since the judgment of two of the raters seems to be opposed to the ranking of the UCI measure. We can hypothesize some possible explanations for this class in particular. As explained in section 6.1, Class 5 represents Facebook posts from pages related to viral content and 'memes'. The poor agreement between human and automated scores could be explained by the lack of textual description on shared pictures and videos, but further exploratory analysis should be made in that area.

7. CONCLUSIONS

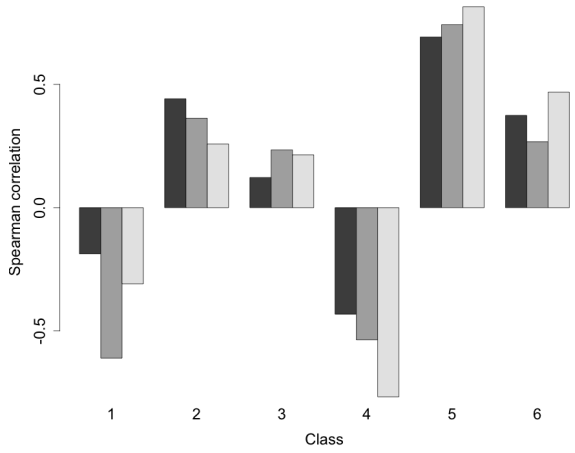

Figure 2: Correlation between UMass and annotators

In this paper, we have compared and contrasted two methods for measuring topic coherence. We tested two measures of automatic evaluation for learned topics against posts collected from Facebook and assessed the compliance of the measures with the human annotators. Analyzing the outcomes of the experiments conducted we conclude that UCI presents good agreement with the human judgment on comprehension for this specific task. This means that external sources like Wikipedia can be used to validate the learned topics. However, further analysis on this matter should be performed in the future in order to materialize this preliminary evidence. In contrast, UMass presented lower values of correlation against the human judgment for this task. This implies that while we have found a good measure for automatically capturing topic comprehension, we were not that successful in identifying a measure for the automatic determination of topic cohesion in such small texts as the ones in Facebook posts.

Another goal of the paper was to test if this kind of evaluation could help end users to filter irrelevant learned topics from a large amount of documents. The experiment demonstrated that UCI, in particular, did a good job when compared with human evaluation. Nevertheless, end users expectations are high, and there is still significant room for improvement regarding modeling human evaluation.

While the methods used in this paper presented interesting results on automatic evaluation for learned topics, the scope of the research can be extended in several other directions in the future. One exciting path is to study how coherence measures could be applied to help summarization methods to produce and evaluate good document summaries. Evaluating not only the coherence among isolated terms, but whole sentences in order to produce complete and coherent texts. Another plan would be to apply a different method to select representative words for each topics using a weighting scheme to either substitute or complement the current approach which is solely based on words frequency.

8. ACKNOWLEDGMENTS

Supported by Integrated project TEC4Growth (including RL1 SMILES, RL2 FourEyes and RL3 iMAN) "Project "TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020" is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF)." and partly financed by Capes/Brazil. Many thanks to the voluntary collaboration by Ciro Oiticica and Lucas Canavarro as annotators of this study.

References

[1] Nikolaos Aletras and Mark Stevenson. "Evaluating Topic Coherence Using Distributional Semantics". In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*. 2013.

[2] Loulwah Alsumait et al. "Topic Significance Ranking of LDA Generative Models". In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*. ECML PKDD '09. Bled, Slovenia: Springer-Verlag, 2009, pp. 67–82.

[3] Anat Ben-David and Ariadna Matamoros Fernández. "Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain". In: *International Journal of Communication* 10 (2016).

[4] W. Lance Bennett and Alexandra Segerberg. *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics*. Cambridge University Press, 2013.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. 1st. O'Reilly Media, Inc., 2009.

[6] David M. Blei. "Probabilistic Topic Models". In: *Commun. ACM* 55.4 (Apr. 2012), pp. 77–84.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022.

[8] Manuel Castells. *Communication Power*. 2009.

[9] Manuel Castells. *Redes de Indignacion y Esperanza*. Alianza Editorial, 2012.

[10] Jonathan Chang et al. "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Neural Information Processing Systems*. Vancouver, British Columbia, 2009.

[11] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. "Termite: Visualization Techniques for Assessing Textual Topic Models". In: *Advanced Visual Interfaces*. 2012.

[12] Javier Toret (coord.) *Tecnopolítica y 15M: La potencia de las multitudes conectadas*. Barcelona, 2013.

[13] Ali Daud et al. "Knowledge discovery through directed probabilistic topic models: a survey". In: *Frontiers of Computer Science in China* 4.2 (2010), pp. 280–301.

[14] Yi Fang et al. "Mining Contrastive Opinions on Political Texts Using Cross-perspective Topic Model". In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. WSDM '12. Seattle, Washington, USA: ACM, 2012, pp. 63–72.

[15] Toshiaki Funatsu et al. “Extracting Representative Words of a Topic Determined by Latent Dirichlet Allocation”. In: *Proc. The Sixth International Conference on Information, Process, and Knowledge Management (eKNOW 2014)*. 2014.

[16] Justin Grimmer. “A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases”. In: *In Proceedings of the First Workshop on Social Media Analytics, SOMA ’10*. 2009.

[17] Katrina Kimport Jennifer Earl. *Digitally Enabled Social Change: Activism in the Internet Age*. MIT Press, 2011.

[18] Bruno Latour. “Beware, your imagination leaves digital traces”. In: *Times Higher Literary Supplemen* (2007).

[19] David Mimno et al. “Optimizing Semantic Coherence in Topic Models”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 262–272.

[20] Jaimie Murdock and Colin Allen. “Visualization Techniques for Topic Model Checking”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI’15. Austin, Texas: AAAI Press, 2015, pp. 4284–4285.

[21] David Newman et al. “Automatic Evaluation of Topic Coherence”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT ’10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 100–108.

[22] David Newman et al. “Evaluating Topic Models for Digital Libraries”. In: *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. JCDL ’10. Gold Coast, Queensland, Australia: ACM, 2010, pp. 215–224.

[23] David Newman et al. “Evaluating Topic Models for Digital Libraries”. In: *Proceedings of the 10th Annual Joint Conference on Digital Libraries*. JCDL ’10. Gold Coast, Queensland, Australia: ACM, 2010, pp. 215–224.

[24] Quentin Pleplé. “Interactive Topic Modeling”. PhD thesis. University of California, San Diego, 2013.

[25] Bernhard Rieder. “Studying Facebook via Data Extraction: The Netvizz Application”. In: *Proceedings of the 5th Annual ACM Web Science Conference*. WebSci ’13. Paris, France: ACM, 2013, pp. 346–355.

[26] Margaret E. Roberts et al. “The Structural Topic Model and Applied Social Science”. In: *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. Peer-Reviewed Conference Workshop. Selected for Oral Presentation. Lake Tahoe, Nevada, 2013.

[27] Carson Sievert and Kenneth E. Shirley. “LDAvis: A method for visualizing and interpreting topics”. In: *ACL Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, 2014.

[28] Hanna M. Wallach et al. “Evaluation Methods for Topic Models”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML ’09. Montreal, Quebec, Canada: ACM, 2009, pp. 1105–1112.

[29] William Warner and Julia Hirschberg. “Detecting Hate Speech on the World Wide Web”. In: *Proceedings of the Second Workshop on Language in Social Media*. LSM ’12. Montreal, Canada: Association for Computational Linguistics, 2012, pp. 19–26.

[30] Wayne Xin Zhao et al. “Comparing Twitter and Traditional Media Using Topic Models”. In: *Proceedings of the 33rd European Conference on Advances in Information Retrieval*. ECIR’11. Dublin, Ireland: Springer-Verlag, 2011, pp. 338–349.