Topic Coherence Metrics: How Sensitive Are They?

João Marcos Campagnolo[1], Denio Duarte[1], Guillherme Dal Bianco[1]

Federal University of Fronteira Sul
Campus Chapecó
Chapecó, Brazil
jota.campagnolo@gmail.com
{duarte,guilherme.dalbianco}@uffs.edu.br

Abstract. Topic modeling approaches extract the most relevant sets of words (grouped into so-called topics) from a document collection. The extracted topics can be used for analyzing the latent semantic structure hiding in the collection. This task is intrinsically unsupervised (without information about the labels), so evaluating the quality of the discovered topics is challenging. To address that, different unsupervised metrics have been proposed, and some of them are close to human perception, *e.g.*, coherence metrics. Moreover, metrics behave differently when facing noise (*i.e.*, unrelated words) in the topics. This article presents an exploratory analysis to evaluate how state-of-the-art metrics are affected by perturbations in the topics. By perturbation, we mean that intruder words are synthetically inserted into the topics to measure the metrics' ability to deal with noises. Our findings highlight the importance of overlooked choices in the metrics sensitiveness context. We show that some topic modeling metrics are highly sensitive to disturbing; others can handle noisy topics with minimal perturbation. As a result, we rank the chosen metrics by sensitiveness, and as the contribution, we believe that the results might be helpful for developers to evaluate the discovered topics better.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: Coherence Metrics, Model Evaluation, Sensibility, Topic Modeling, Unsupervised Machine Learning

1. INTRODUCTION

Probabilistic topic modeling approaches aim to extract sets of words from a document collection. Those sets arrange the documents into topics [Steyvers and Griffiths 2007; Blei 2012]. We can use the discovered topics to analyze the latent semantic structure in a given collection [O'Callaghan et al. 2015]. One way of quantifying the discovered topics' coherence is to evaluate the topics' effectiveness in a given application domain. Coherence metrics have been proposed to accomplish the evaluation. Nevertheless, evaluating the coherence of the discovered topics automatically, as an unsupervised task, gives no guarantee on the topic model interpretability.

Since its inception, automatic evaluation of topic quality has been a challenge. The problem is that topics are more natural to be assessed by humans, and automatic metrics cannot capture human interpretability [Nikolenko 2016]. Röder et al. [2015] proposed a unifying framework for quantifying coherence metrics. They conducted an extensive study of coherence metrics performance and human evaluation on topics seen as sets of words. Other works propose similar evaluations: measuring the human-interpretability of topics, for example, Chang et al. [2009] and Lau et al. [2014]. Those works attempt to check whether or not the discovered topics are coherent regarding the metrics and the human interpretability. Although several works evaluate topic model approaches and human interpretability, such works do not evaluate the metrics' behavior against the topic purity, that is, the range of results facing topics with very related words and topics with intruder words. Understanding the behavior of metrics is essential to machine learning developers because a given metric works differently regarding the model and dataset. For example, in regression problems, if the developer wants to penalize large errors, they apply Root Mean Squared Error (RMSE) metric; otherwise, Mean Absolute Error (MAE) is a better choice because it is more robust to data with outliers. This shows the importance of choosing a specific metric to evaluate a given machine learning model. Moreover, there is no study to analyze the metrics' behavior concerning the discovered topics' quality, to the best of our knowledge.

In this work, we propose an evaluation of the metrics applied in topic modeling regarding their sensitiveness. The sensitiveness is measured by observing the metrics' behavior against good and noisy topics, that is, the variation they display in different topic scenarios. To accomplish that, we produce two types of topics: pure (composed of only related words) and noisy (pure topics with intruded words). We validate the topics through a survey with over 60 respondents. The produced topics are about well-known subjects: politics, religion, music, and Christmas. The chosen coherence metrics are those presented in the state of the art of topic modeling [Röder et al. 2015]: C_V, C_P, C_{UCI}, C_{UMass}, C_{NPMI}, and C_A.

The patterns against pure and noisy topics are compared to identify the metric sensitiveness. It allows us to determine that some metrics are most sensitive (in all topics analyzed) and should be used when pure topics must be achieved. However, when noise is not a drawback, some metrics show more potential to be applied. For example, metrics that behave close to the human evaluation (see [Röder et al. 2015] for further details). The main contribution of this work is to rank the state-of-the-art coherence metrics regarding their behavior over different topics: from pure ones, *i.e.*, topics composed of only related words, to very noisy ones, *i.e.*, topics composed of only unrelated words. This behavior allows us to classify those metrics according to noise sensitiveness. Note that we do not intend to highlight the best coherence metric; instead, compare them regarding how well they deal with noise.

The rest of this work is organized as follows. The following section reviews the coherence metrics used in our exploratory analysis. Section 3 briefly presents the related work. The methodology used to build the topics for the exploratory analysis is presented in Section 4. Section 5 presents and discusses the results of the investigation. Finally, Section 6 concludes this work.

2. TOPIC COHERENCE METRICS

This section presents briefly the metrics used in this study. For a more detailed discussion, we refer readers to the works [Bouma 2009; Aletras and Stevenson 2013; Röder et al. 2015].

The models generated by topic modeling approaches are hard to evaluate since unsupervised learning algorithms do not have labels that verify the correctness of the obtained results [Chang et al. 2009; Lau et al. 2014]. The best way to evaluate unsupervised models is by using human evaluation; however, this evaluation can become costly for large volumes of data.

Metrics based on coherence capture the co-occurrence frequencies of terms within a reference corpus and distributional semantics. The intuition is that terms co-occurring frequently or close to each other within a semantic space are likely to contribute to higher coherence levels. The measure, thus, relies on the top words of a topic and co-occurrence counts gathered from the corpus.

To present the metrics properly, we first introduce the segmentation of word subsets. The segmentation aims at building a set of pairs of words from a given document. The segmentation helps to identify how the words appear together, *i.e.*, next to or far from each other regarding the segmentation. In the following, we present two strategies to segment words.

Sliding Window: a *sliding window* is a subset of consecutive words of size N that can be moved word by word to either side. For example, Figure 1 shows three different possibilities of a size four sliding window in a given document: sw_y, sw_g, and sw_r.

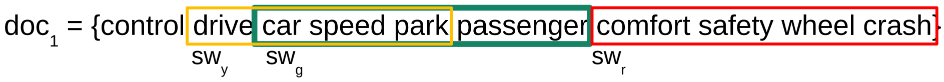

doc₁ = {control drive car speed park passenger comfort safety wheel crash}
sw y
sw g
sw r

Fig. 1. Sliding windows examples with size four.

Context Window: a *context window* is a subset of N (N defines the window size and it is greater than 0) consecutive words located immediately next to a given word. Figure 2 shows the same document as Figure 1 using a context window of size three around the word *park*. In this case, *park* is compared to words *drive*, *car*, and *speed* on its left side, and words *passenger*, *comfort*, and *safety*, on its right.

The probabilities of sliding are estimated using the word segmentation. The documents are seen as a copy of the window content where the word counts are determined. After the probability calculation, confirmation measures are applied. The confirmation measures compute how strong a conditioning word W^* set supports another word set W'. For example, the difference-measure is calculated as follows:

$$(W', W^*) = P(W' \mid W^*) - P(W')$$

Before presenting the metrics used in this work, we present the confirmation measures used by the metrics.

Pointwise Mutual Information (PMI): PMI is used to measure the associativity between two words. PMI is calculated from a word occurrence count and is given by the following equation:

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)}\right)$$

Given the equation, $P(w_i, w_j)$ is the frequency/probability of observing the words w_i and w_j in the same window (*context* or *sliding*). $P(w_i)$ and $P(w_j)$ are, respectively, the frequency/probability of observing the words w_i and w_j separately. The closer the frequency of co-occurrence of two words is to the occurrence of the two words separately, the better the score for the given pair of words. The ϵ constant can be used to prevent the occurrence of a zero logarithm.

Normalized Pointwise Mutual Information (NPMI): NMPI is a variation of *PMI* that normalizes the value obtained to the interval $[-1, 1]$. The lower limit -1 means no co-occurrence, 0 means independence between the two words, and 1 means complete co-occurrence. The formula for *NPMI* is given by:

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log\left(P(w_i, w_j)\right)}$$

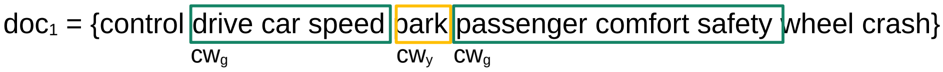

doc₁ = {control drive car speed park passenger comfort safety wheel crash}
cwg
cwy
cwg

Fig. 2. Example of size three context window for word park.

Confirmation Measure of Fitelson's Coherence: denoted by m_f, Fitelson [Fitelson 2003] proposed this function to rate the relationship between two propositions. This function is also used by Röder et al. [2015] to rate the relationship between a word w_i and a subset of words $S(i)_j$, and is defined by the following formula:

$$m_f(w_i, S(i)_j) = \frac{P(W_i|S(i)_j) - P(W_i|\neg S(i)_j)}{P(W_i|S(i)_j) + P(W_i|\neg S(i)_j)}$$

For all metrics below, we use the topic $t_1 = \{\mathit{car}, \mathit{driver}, \mathit{wheel}, \mathit{speed}\}$ as a running example (when it applies). As we used the Palmetto framework to run the experiments, all parameters (*e.g.*, the size of sliding windows) for the metrics are pre-defined. We refer readers to [Röder et al. 2015] for further details.

2.1 UMass Coherence

UMass Coherence (C_{UMass}) is a specialization of the metric proposed by [Mimno et al. 2011], and it is based on the equation:

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \left(\frac{P(w_i, w_j) + \epsilon}{P(w_j)} \right)$$

Given topic t_1, the calculation of C_{UMass} is:

$$\begin{aligned} C_{UMass}(t_1) = \frac{2}{4 \cdot (4-1)} &\cdot (\log (P(\mathit{driver}|\mathit{car})) + \log (P(\mathit{wheel}|\mathit{car})) + \\ &+ \log (P(\mathit{wheel}|\mathit{driver})) + \log (P(\mathit{speed}|\mathit{car})) + \\ &+ \log (P(\mathit{speed}|\mathit{driver})) + \log (P(\mathit{speed}|\mathit{wheel}))) \end{aligned}$$

Generally, word probabilities are calculated according to their occurrence in documents on a given collection.

2.2 UCI Coherence

The *UCI Coherence* metric (C_{UCI}) is based on a *sliding window* with size 10 and the *PMI* of all word pairs of a topic *N-top* words, defined by the following formula:

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j)$$

The *PMI* is calculated for each pair of words, and the arithmetic mean of these values results from C_{UCI}. For example, $C_{UCI}(t_1)$ is:

$$\begin{aligned} C_{UCI}(t_1) = \frac{2}{4 \cdot (4-1)} &\cdot (PMI(\mathit{car}, \mathit{driver}) + PMI(\mathit{car}, \mathit{wheel}) + \\ &+ PMI(\mathit{car}, \mathit{speed}) + PMI(\mathit{driver}, \mathit{wheel}) + \\ &+ PMI(\mathit{driver}, \mathit{speed}) + PMI(\mathit{wheel}, \mathit{speed})) \end{aligned}$$

2.3 NPMI Coherence

The *NPMI Coherence* metric (C_{NPMI}) is an enhanced version of C_{UCI}, which uses *NPMI* instead of *PMI*, and is defined by the following formula:

$$C_{NPMI} = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} NPMI(w_i, w_j)$$

Using the topic t_1 as example, $C_{NPMI}(t_1)$ is:

$$\begin{aligned} C_{NPMI}(t_1) = \frac{2}{4 \cdot (4-1)} \cdot (&NPMI(car, driver) + NPMI(car, wheel) + \\ &+ NPMI(car, speed) + NPMI(driver, wheel) + \\ &+ NPMI(driver, speed) + NPMI(wheel, speed)) \end{aligned}$$

2.4 C_P Coherence

C_P evaluates the coherence of a topic using the *Confirmation Measure of Fitelson's Coherence* from a *sliding window* with size 70 based on the following equation:

$$C_P = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=2}^{N} \sum_{j=1}^{i-1} m_f(w_i, w_j)$$

where m_f is the Confirmation Measure of Fitelson's Coherence between the current word and the previous one ranking by importance (*i.e.*, the probability of belonging to the given topic). For example, using t_1, the result is:

$$\begin{aligned} C_P(t_1) = \frac{2}{4 \cdot (4-1)} \cdot (&m_f(driver, car) + m_f(wheel, car) + \\ &+ m_f(wheel, driver) + m_f(speed, car) + \\ &+ m_f(speed, driver) + m_f(speed, wheel)) \end{aligned}$$

2.5 C_A Coherence

C_V measures the combination of all word pairs in a given topic using a variation of *NPMI* (as C_V does) from a context window with size 5. Given two words *car* and *drive*, C_A is calculated as follows:

$$v_{\text{car,driver}} = NPMI(car, driver)^{\gamma} = \left(\frac{PMI(car, driver)}{-\log\left(P(car, driver) + \epsilon\right)} \right)^{\gamma}$$

The above calculation results in a set of vectors that are used to find the coherence. Based on topic t_1:

$$\begin{aligned} \vec{v}_{\text{car}} = \{&NPMI(car, car)^{\gamma}, NPMI(car, driver)^{\gamma}, \\ &NPMI(car, wheel)^{\gamma}, NPMI(car, speed)^{\gamma}\} \end{aligned}$$

Finally, the vector distance is measured using cosine similarity[1]:

$$C_A = \frac{1}{6} \cdot ((\cos(\vec{v}_{\text{car}}, \vec{v}_{\text{driver}}) + \cos(\vec{v}_{\text{car}}, \vec{v}_{\text{wheel}}) + \cos(\vec{v}_{\text{car}}, \vec{v}_{\text{speed}}) + \\ + \cos(\vec{v}_{\text{driver}}, \vec{v}_{\text{wheel}}) + \cos(\vec{v}_{\text{driver}}, \vec{v}_{\text{speed}}) + \cos(\vec{v}_{\text{wheel}}, \vec{v}_{\text{speed}}))$$

2.6 C_V Coherence

C_V uses a variation of *NPMI* to calculate the coherence over a sliding window with size 110. It calculates the co-occurrence of a word of a given topic against all words of the same topic. The weight γ is used to give more strength to more associative words. Using t_1 and the words *car* and *driver*, a vector for these two words is created as follows:

$$v_{\text{car,driver}} = NPMI(car, driver)^{\gamma} = \left(\frac{PMI(car, driver)}{-\log(P(car, driver) + \epsilon)} \right)^{\gamma}$$

Secondly, a vector for every word is created as follows (in the example, the *car* vector is created):

$$\vec{v}_{\text{car}} = \{NPMI(car, car)^{\gamma}, NPMI(car, driver)^{\gamma}, \\ NPMI(car, wheel)^{\gamma}, NPMI(car, speed)^{\gamma}\}$$

Finally, the vector distances are measured using cosine similarity:

$$\vec{v}_c = \vec{v}_{\text{car}} + \vec{v}_{\text{driver}} + \vec{v}_{\text{wheel}} + \vec{v}_{\text{speed}}$$

$$C_V = \frac{1}{4} \cdot (\cos(\vec{v}_{\text{car}}, \vec{v}_c) + \cos(\vec{v}_{\text{driver}}, \vec{v}_c) + \cos(\vec{v}_{\text{wheel}}, \vec{v}_c) + \cos(\vec{v}_{\text{speed}}, \vec{v}_c))$$

Note that C_V and C_A are very similar, but the size of sliding window.

Note that for all the metrics, the higher the value, the better performance.

3. RELATED WORK

Metrics are fundamental to enable direct measurement between models and evaluate their performance. In supervised learning, for example, in which the input datasets are equipped with the label information, the evaluation process is straightforward. Metrics, such as accuracy, precision, recall, and Fβ-score, can be applied to match the predicted class with the ground-truth classes [Alvarez 2002; Fatourechi et al. 2008; Folleco et al. 2008; Powers 2011; Juba and Le 2019]. On the other hand, unsupervised metrics are a little trickier since there is no label to compare the ground-truth and the predicted values [Duarte and Ståhl 2019].

Topic modeling approaches are meant to discover topics from document collections regarding the frequency of the words. Based on the frequency, latent topics could be revealed – the topics emerge from the analysis of the original documents [Blei 2012]. This is an unsupervised learning task since there are no labels to check discovered topics against gold-standard topics. Accordingly, topic modeling

[1] It is a measurement that quantifies the similarity between two or more vectors. The similarity is based on the cosine of the angle between vectors [Vijaymeena and Kavitha 2016].

metrics play a crucial role in evaluating the discovered topics. Most of the topic modeling metrics work focuses on assessing modeling approaches or evaluating the topic's human interpretability against the metrics results.

In [Röder et al. 2015], a set of coherence metrics performance is compared with human evaluation in the topic modeling context. A proposed framework compares seven metrics to the ranking induced by human ratings. They grouped the coherence measure into two groups: direct and indirect measures. The former computes the confirmation of a single pair of words, and the latter computes the confirmation between a word and a set of words. They identify that C_V is the best indirect coherence measure related to the human rating, and C_P is the best direct one.

Like Chang et al. [2009] and Lau et al. [2014], other works compare metrics to evaluate the quality of the topics inferred by the model and how well the model assigns topics to documents. Models are built from document collections using different approaches (*e.g.*, probabilistic Latent Semantic Indexing – pLSI, Latent Dirichlet Allocation – LDA, and Correlated Topic Model – CTM) and compared the discovered topics to human evaluation on the same topics to accomplish that evaluation. The topics produced by the models are mixed up with word intrusion, and humans have to identify the intruder. For evaluating a set of discovered topics, an intruder topic is inserted in the set. Again, humans have to identify the intruder. The authors conclude that likelihood-based measures are not suitable to evaluate topics. Newman et al. [2010] also assess metrics that measure topic coherence based on human rating. Top-n words represent topics, and humans are asked to rate the topics as good, neutral, or bad. However, as the previous work, it aims to identify which metric is closer to human evaluation.

Unlike the works previously presented, our work does not compare metrics to human evaluation. We explore the behavior of the metrics concerning the purity of the topics. Instead of showing the best metrics regarding human interpretability, we intend to measure the metrics' behavior on different types of topics.

4. METHODOLOGY TO BUILD THE TOPICS

This section presents the intuition behind the construction of pure and noisy topics. We show how non-expert users are employed to validate the topics.

The experiments were based on five defined subjects reflecting well-known domains, *i.e.*, do not require specific technical knowledge to understand them. The subjects are *sports*, *politics*, *religion*, *music*, and *Christmas*. The definition of the topics' subjects was followed by the choice of the words that compose each topic, validation of the built sets, and finally, the application of coherence metrics. The sports subject was used as a pilot through this methodology.

4.1 Building the Topics

For each subject, the ten most-accessed news articles about the subject were extracted from The *New York Times* website and the English version of *Wikipedia*. Accordingly, five collections were built, one for each subject. The collections were pre-processed by removing the stop words and special symbols and applying tokenization and stemming techniques. After the pre-processing steps, we ordered the words by occurrence.

We picked the top-10 nouns from the ranking to build the topics. For example, to compose the topic about sports, the words *player* and *game* were selected, but the words *said* and *open* were ignored.

We built two sets of words for each subject: one with five words (5-word topic) and another with ten words (10-word topic). The latter is the union of the 5-word topic with the following five better-ranked words. We refer to $Subject_5$ for the $Subject$'s 5-word and $Subject_{10}$ for the $Subject$'s 10-word. Table I shows the top-10 words for the proposed topics (subjects). For the *Sports* subject, for example,

Table I. Selected words by topics subject.

Sports$_{10}$	**Politics**$_{10}$	**Religion**$_{10}$	**Music**$_{10}$	**Christmas**$_{10}$
player	govern	belief	play	celebration
game	president	church	song	holiday
match	election	moral	instrument	gift
competition	public	god	sound	santa
team	constitution	tradition	composition	december
tournament	corruption	faith	melody	decoration
league	congress	spiritual	concert	tree
athlete	campaign	sacred	singer	birth
score	senator	holy	symphony	jesus
goal	law	supernatural	rhythm	feast

Table II. Intrusive words at topics with *size=5*.

Original	**1 Intruder**	**2 Intruders**
word1	word1	word1
word2	word2	word2
word3	word3	word3
word4	word4	***wordA1***
word5	***wordA1***	***wordO1***

the 5-word topic is $Sports_5 = \{player, game, match, competition, team\}$ and the 10-word topic is $Sports_{10} = Sports_5 \cup \{ tournament, league, athlete, score, goal\}$.

The above steps built two sets of words for each subject, and we call them pure topics because the words are significantly related to the subject. The following steps are to build new noisy topics, that is, topics including the intruding words.

We use the word intrusion technique [Chang et al. 2009] to create noisy topics for each subject. This technique replaces words from a topic with words that are not associated with it. Those words are easily identified as intruders. As an example, given two topic $T_1 = \{dog, cat, horse, pig, cow\}$ and $T_2 = \{dog, cat, car, pig, cow\}$. It is easy to identify *car* as the intruder word in T_2. Note that *horse* is replaced with *car* in T_1. Supposing that the topics are about *Animals*, clearly, *car* does not belong to this subject.

We built two noisy topics for each 5-word pure topic: one with one intruder word and another with two intruder words. Table II shows the intuition behind the construction of the two noisy topics.

The construction is as follows:

—First intruding word (*wordA1*): we randomly pick a noun from the set of words that occur within the average occurrence in the collection, that is, a word slightly related to the topic.

—Second intruding word (*wordO1*): we randomly pick a noun from the set of words that occur only once in the collection, that is, a word that is very unrelated to the topic, even if it appears in the collection.

The 10-word noisy topics were built using the same previous reasoning. However, four new noisy topics were proposed: (*i*) nine words and $wordA1$, (*ii*) eight words and $wordA1$ $wordO1$, (*iii*) seven words and $wordA1$ $wordO1$ $wordA2$, and (*iv*) six words and $wordA1$ $wordO1$ $wordA2$ $word02$. We refer to the pure topics as $Subject_{n-0}$, where n is 5 or 10, and the noisy topics as $Subject_{n-m}$, where m is the number of intruding words.

In the end, we have 15 5-word topics and 25 10-word topics – Table III and Table IV present the $Sports_{5-m}$ and $Sports_{10-m}$ topics, respectively, built from the previous steps. The intruding words

Table III. Generated topics with *size=5* for *Sports* subject.

$Sports_{5-0}$	$Sports_{5-1}$	$Sports_{5-2}$
player	player	player
game	game	game
match	match	match
competition	competition	***watch***
team	***watch***	***tree***

Table IV. Generated topics with *size=10* for *Sports* subject.

$Sports_{10-0}$	$Sports_{10-1}$	$Sports_{10-2}$	$Sports_{10-3}$	$Sports_{10-4}$
player	player	player	player	player
game	game	game	game	game
match	match	match	match	match
competition	competition	competition	competition	competition
team	team	team	team	team
tournament	tournament	tournament	tournament	tournament
league	league	league	league	***watch***
athlete	athlete	athlete	***watch***	***tree***
score	score	***watch***	***tree***	***push***
goal	***watch***	***tree***	***push***	***mirror***

are in boldface. Note that the words *watch* and *push* are chosen as slightly related to the topic and the words *tree* and *mirror* as the unrelated ones in $Sports_{10-4}$.

We also built two very noisy topics: one with five words ($Noisy_5$) and another with ten words ($Noisy_{10}$). The five words are picked randomly from the proposed topics to comprise $Noisy_5$ = {*team*, *govern*, *belief*, *melody*, *feast*}. As we could not take more than five words from the proposed topics, we add five new unrelated words to $Noisy_5$ to build $Noisy_{10}$, *i.e.*, $Noisy_{10}$= $Noisy_5 \cup$ {*cyanic*, *satellite*, *oxter*, *cybersquatting*, *canorous*}. Humans validated all the proposed topics through a survey, but the noisy ones.

4.2 Topics Validation

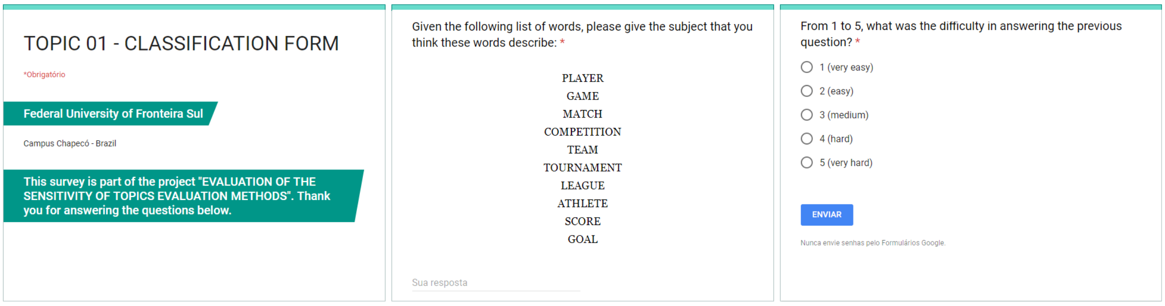
Here we present the methodology used to validate the proposed topics. In total, 63 respondents were selected to answer the surveys, separated into seven distinct groups. Each group responded only to a specific type of form. To prevent responses from being influenced by the form of the other groups, none of the respondents had access to other groups' forms. The seven forms (one for each group) are one to validate the pure topics, two to validate the intruder words for 5-word topics, and four to validate the intruder words for 10-word topics.

All the responses were collected anonymously through the *Google Forms* tool. We built two forms: validating the pure topics and checking whether or not the respondents recognize the intruder words. The respondents' profile is Ph.D. and students from Sweden and Brazil, undergraduate students from Brazil, and Brazilian researchers.

Figure 3 shows an example of one form to check if the chosen words for $Sports_{10-0}$ lead the respondents to choose sport as a subject. Note that the questionnaire survey for pure topics is composed of one open answer, and the Likert Scale Question [Likert 1932] is used to verify how hard it was to answer the open question (1 (very easy) to 5 (very hard)).

The form for noisy topics is slightly different: the respondent must select the intruding word(s) (checkbox) and answer a question on the level of difficulty to find the intruding word(s). For finding the intruder words, the respondents knew the topic subject.

We first used the *Sports* subject as a pilot to identify any possible problems during the validation.



TOPIC 01 - CLASSIFICATION FORM

*Obrigatório

Federal University of Fronteira Sul

Campus Chapecó - Brazil

This survey is part of the project "EVALUATION OF THE SENSITIVITY OF TOPICS EVALUATION METHODS". Thank you for answering the questions below.

Given the following list of words, please give the subject that you think these words describe: *

PLAYER
GAME
MATCH
COMPETITION
TEAM
TOURNAMENT
LEAGUE
ATHLETE
SCORE
GOAL

Sua resposta

From 1 to 5, what was the difficulty in answering the previous question? *

○ 1 (very easy)
○ 2 (easy)
○ 3 (medium)
○ 4 (hard)
○ 5 (very hard)

ENVIAR

Nunca envie senhas pelo Formulários Google.

Fig. 3. Forms to check the validity of a proposed topic.

Table V. Results for the pure topics.

			Difficulty				
Topic	**Answers**	**# of R**	**VE**	**E**	**M**	**H**	**VH**
Sports	game, sport, soccer	8	4	3	1	0	0
Politics	politic, politician, comunism	5	2	2	0	1	0
Religion	Religion, religiosity	5	2	2	0	1	0
Music	music, band	5	2	2	0	1	0
Christmas	christmas	5	2	2	0	1	0

Table VI. Answers for noisy topics with one intruding word.

				Difficulty				
Topics	**Words**	**Chosen ones**	**# of R**	**VE**	**E**	**M**	**H**	**VH**
Sports	player game match competition **watch**	watch(11), game(1), competition(1)	13	0	6	5	1	1
Politics	election **media** public govern president	media(6), public(1)	7	1	3	3	0	0
Religion	moral belief **trade** church god	trade	7	1	3	3	0	0
Music	play song **contrast** instrument sound	contrast	7	1	3	3	0	0
Christmas	holiday santa gift celebration **case**	case	7	1	3	3	0	0

The pilot step performed as expected; thus, the other topics were sent to be validated. Table V shows the results for the pure topics. All the subjects were successfully identified, and the difficulty of most of them was *easy*. For example, five respondents (Column *# of R*) answered the Topic Music's words as *music* or *band*. Four of them chose *very easy* (VE) or *easy* (E) to identify the subject, and one chose *hard* (H). Based on the answers, we can conclude that the proposed pure topics are suitable for our experiments.

Table VI presents the answers for one intruding word in 5-word topics. Note that the hardest intruding word to find was related to the Sports subject. The overall results showed that eleven over thirteen (85%) of respondents found the right intruder. An example of a question in the form is: *"Given the following list of words, four are about politics, and one is not. Identify which word is not about politics:"*. Note that the Sports topics have more reviews since the subject was used as a pilot for our methodology.

The other non-pure topics (the noisy ones) validation followed the same results as previously presented. At the end of the survey, we have 45 topics validated by humans ready to be used in our experiments.

4.3 Threats to Validity

We propose topics by extracting words from two collections: The New York Times and Wikipedia. This step helped us find the *right* words for the subjects/topics presented. However, no matter how the words were chosen, *e.g.*, querying a collection or using a topic modeling approach, we must confirm that the topics are valid regarding human perception. After conducting a survey, we have this confirmation.

We selected non-native English speakers who have a good command of English as respondents. We claimed that the major limitation of the proposed topics, and therefore a threat to its validity, refers to mistakes that respondents can make when answering the survey. We mitigated this threat by presenting the research and answering, if necessary, any questions about the survey. We highlight that the authors chose the respondents, and so, they were not randomly picked. All respondents are from academia and, as stated before, knew the topics, as the proposed subjects are known worldwide. Moreover, we have at least seven respondents for every proposed noisy topic, which allows us to choose the most chosen option. The right intruder word was the most chosen in all cases, and we did not have to reapply for the survey (see Table VI).

5. METRICS SENSITIVITY

This section presents the six state-of-the-art metrics' sensitivity, *i.e.*, C_V, C_P, C_{UCI}, C_{UMass}, C_{NPMI}, and C_A. We use the Palmetto framework, which implements these metrics and uses three million English Wikipedia articles as an external corpus [Röder et al. 2015]. We check the sensitivity by analyzing the behavior of each metric against the proposed topics. Moreover, we expect the metrics' performance to worsen progressively due to the topics' noise (*i.e.*, the intruder words).

$Subject_{n-0}$ and $Noisy_n$ (where n is defined as 5 or 10) are used as our parameters for sensitiveness. $Subject_{n-0}$ represents the upper limit, *i.e.*, the best result for a given metric, and $Noisy_n$, the lower limit, *i.e.*, the worst result. Therefore, the results for $Subject_{n-m}$ $(m > 0)$ are placed between the limits.

C_{UMass} and C_A are not symmetric since the word order changes the result. To avoid the resulting bias, we applied these two metrics changing the position of all topic words and calculating their mean.

5.1 5-words Topics

Table VII presents the performance of all metrics regarding topics composed of five words. Notice that the absolute values outputted by the metrics are on a different scale. For example, C$_{UMass}$ outputs negative values. Note also that topics about Sport get the best metrics' performance. We believe that it is due to sports being a more popular subject. On the other hand, topics about Christmas get worse results. That can indicate Christmas topics are less popular than the others. The topics' coherence is checked against Wikipedia collection. So, there are more documents containing words about sports, for example, than about Christmas.

We scaled the metrics outputs to normalized the results to enable comparing the metrics more easily. The scaling step is applied for each topic subject considering only a given metric as follows: (i) we calculated the mean for every metric using $Subject_{n-k}$, where n is equal to 5 and $0 \leq k \leq 2$, and (ii) the means were scaled based on the following equation:

$$x_{new} = \frac{x - min(X)}{max(X) - min(X)}$$

Table VII. Metrics performance in all 5-words topics.

Topics	$\mathbf{C}_V$	$\mathbf{C}_P$	$\mathbf{C}_{UCI}$	$\mathbf{C}_{UMass}$	$\mathbf{C}_{NPMI}$	$\mathbf{C}_A$
Sports$_{5-0}$	0.6103	0.6572	1.3386	-1.4890	0.1583	0.2969
Sports$_{5-1}$	0.5288	0.4117	0.7504	-1.9735	0.0822	0.1922
Sports$_{5-2}$	0.5184	0.1442	0.1705	-2.4105	0.0323	0.1619
Politics$_{5-0}$	0.5163	0.5410	1.0651	-1.3891	0.0976	0.2562
Politics$_{5-1}$	0.4816	0.3117	0.4114	-1.5918	0.0412	0.2019
Politics$_{5-2}$	0.4772	0.0336	-0.3013	-2.4754	-0.0107	0.1860
Religion$_{5-0}$	0.5173	0.6211	1.6398	-1.8808	0.1446	0.2729
Religion$_{5-1}$	0.4988	0.3003	0.6599	-2.1307	0.0647	0.2129
Religion$_{5-2}$	0.4677	0.0694	-0.1440	-2.2613	-0.0059	0.2133
Music$_{5-0}$	0.5159	0.5434	1.1646	-2.1763	0.1139	0.2103
Music$_{5-1}$	0.4937	0.3717	0.7987	-2.3344	0.0800	0.1796
Music$_{5-2}$	0.4666	0.0383	-0.1056	-2.8854	0.0008	0.1095
Christmas$_{5-0}$	0.4713	0.3688	1.0077	-2.4593	0.0804	0.1557
Christmas$_{5-1}$	0.4720	0.2224	0.6071	-2.5607	0.0494	0.1515
Christmas$_{5-2}$	0.4674	0.0921	-1.0140	-2.7023	-0.0227	0.1679
Noisy$_5$	0.4620	-0.2156	-5.2580	-4.1590	-0.2078	0.0913

Table VIII. Average metrics performance for 5-words topics and their scaled values. The values are in the format: $avg \pm std(scld)$ - average (avg), standard deviation (std), and scaled ($scld$).

Topic	$\mathbf{C}_V$	$\mathbf{C}_P$	$\mathbf{C}_{UCI}$	$\mathbf{C}_{UMass}$	$\mathbf{C}_{NPMI}$	$\mathbf{C}_A$
Subj$_{5-0}$	0.53±0.05(1.00)	0.55±0.11(1.00)	1.24±0.25(1.00)	-1.88±0.45(1.00)	0.12±0.03(1.00)	0.24±0.06(1.00)
Subj$_{5-1}$	0.49±0.02(0.51)	0.32±0.07(0.71)	0.65±0.15(0.91)	-2.12±0.37(0.90)	0.06±0.02(0.83)	0.19±0.02(0.65)
Subj$_{5-2}$	0.48±0.02(0.27)	0.08±0.05(0.38)	-0.28±0.44(0.77)	-2.55±0.25(0.71)	0.00±0.00(0.63)	0.17±0.04(0.52)
Noisy$_5$	0.46±0.00(0.00)	-0.22±0.00(0.00)	-5.26±0.00(0.00)	-4.16±0.00(0.00)	-0.21±0.00(0.00)	0.09±0.00(0.00)

Figure 4 plots the metrics' behavior using the values from Table VIII. Firstly, note that $Subject_{5-0}$ and $Noisy_5$ represent the upper and lower limits of the metrics' performance, respectively. Therefore, the line between both represents how the metric reacts against the topics with intruder words. We can identify the behavior of the metrics based on the lines' path from the upper to the lower limits. The analysis is as follows:

—C_{UCI}, C_{UMass}, and C_{NPMI} show similar behavior, and the results over topics with intruder words are above 0.65. For example, C_{UCI}'s worst average result is 0.76 (topic $Subj_{5-2}$). The average variation from topic $Subj_{5-0}$ to topic $Subj_{5-1}$ is about 12%, from topic $Subj_{5-1}$ to topic $Subj_{5-2}$ is about 18%, and from topic $Subj_{5-0}$ to topic $Subj_{5-2}$ is, naturally, about 30%. It shows that these metrics are less sensitive than C_V, C_P, and C_A.

—The two most sensitive metrics are C_V and C_P. They vary on average 33% from $Subj_{5-0}$ to topic $Subj_{5-1}$ and from $Subj_{5-1}$ to topic $Subj_{5-2}$. Moreover, the average variation from $Subj_{5-0}$ to topic $Subj_{5-2}$ is 66%. It shows that difference is double regarding the less sensitive metrics.

—C_A follows C_V and C_P's behavior, but the variation between $Subj_{5-1}$ to topic $Subj_{5-2}$ is less than $Subj_{5-0}$ to topic $Subj_{5-1}$. Moreover, some noisy topics have a better performance than the pure ones. Christmas's topics show this behavior (see Table VII).

5.2 10-words Topics

The analysis of 10-word topics follows the same reasoning used previously since the metrics' results follow similar patterns. Table IX shows the results of every metric against the proposed topics, and Table X presents the result of applying the scaling approach as applied to 5-words topics. Interestingly, C_{UMass} shows inconsistent behavior respecting the pure and one intruder word topics. The average result of the pure topic is worse than the one-word intruder topic. This behavior of C_{UMass} is provoked mainly by the topic of religion (see Table IX). The standard deviation shows that: the pure topic

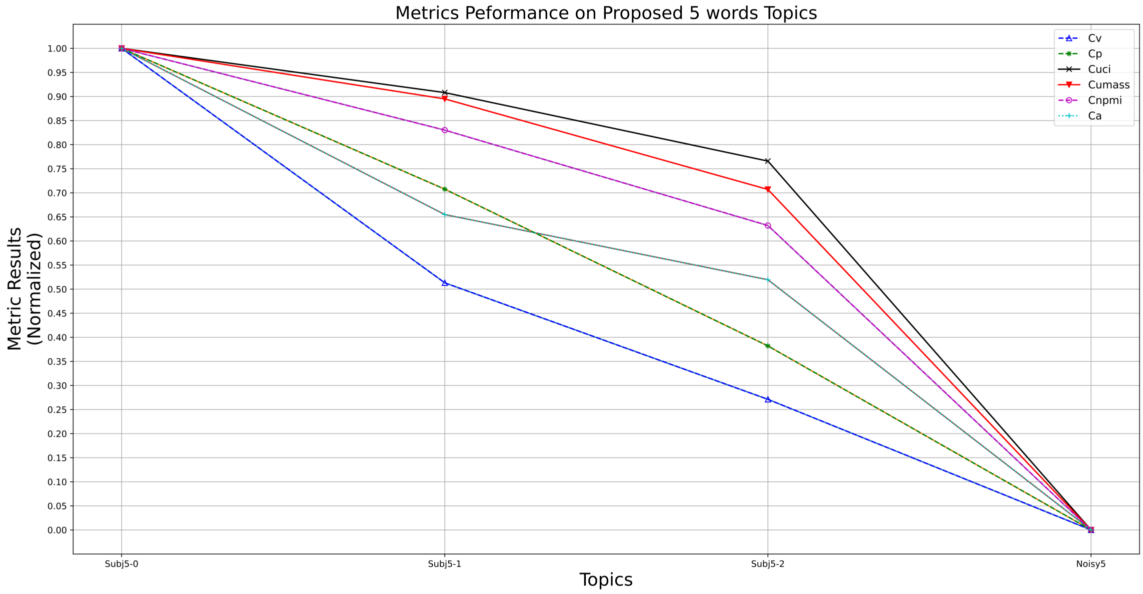


Fig. 4. The average performance of the metrics on the proposed five words topics.

Table IX. Metrics performance in all 10-words topics.

Topics	$\mathbf{C}_{V}$	$\mathbf{C}_{P}$	$\mathbf{C}_{UCI}$	$\mathbf{C}_{UMass}$	$\mathbf{C}_{NPMI}$	$\mathbf{C}_{A}$
Sports$_{10-0}$	0.5683	0.6782	1.4720	-1.7515	0.1687	0.3444
Sports$_{10-1}$	0.5192	0.5515	1.0922	-1.9709	0.1248	0.2744
Sports$_{10-2}$	0.4803	0.3362	0.4724	-2.3238	0.0686	0.2244
Sports$_{10-3}$	0.4662	0.2639	0.3769	-2.3300	0.0590	0.2045
Sports$_{10-4}$	0.4211	0.1605	0.1470	-2.6944	0.0315	0.1588
Politics$_{10-0}$	0.4210	0.5386	1.1254	-1.9009	0.1036	0.2690
Politics$_{1}{}_{0-1}$	0.4043	0.4484	0.8734	-2.1438	0.0805	0.2323
Politics$_{10-2}$	0.3831	0.2722	0.1487	-2.3565	0.0355	0.2065
Politics$_{10-3}$	0.3725	0.2253	-0.0208	-2.3913	0.0197	0.1788
Politics$_{10-4}$	0.3571	0.0909	-0.8423	-2.7631	-0.0228	0.1752
Religion$_{10-0}$	0.4557	0.6889	1.9673	-2.3292	0.1662	0.2944
Religion$_{10-1}$	0.4430	0.5147	1.3921	-2.0406	0.1245	0.2744
Religion$_{10-2}$	0.4141	0.3674	0.9360	-2.0637	0.0852	0.2485
Religion$_{10-3}$	0.3963	0.2478	0.5483	-2.0266	0.0540	0.2110
Religion$_{10-4}$	0.3733	0.0919	0.0585	-2.2775	0.0151	0.1871
Music$_{10-0}$	0.4582	0.6370	1.4056	-2.4042	0.1247	0.2142
Music$_{1}{}_{0-1}$	0.4305	0.5167	1.0827	-2.4747	0.0974	0.1805
Music$_{10-2}$	0.4073	0.3260	0.6082	-2.5147	0.0606	0.1611
Music$_{10-3}$	0.3878	0.2345	0.4666	-2.6016	0.0468	0.1479
Music$_{10-4}$	0.3725	0.1405	0.0029	-2.8473	0.0203	0.1340
Christmas$_{10-0}$	0.3527	0.3351	0.8332	-2.8237	0.0649	0.1619
Christmas$_{1}{}_{0-1}$	0.3346	0.1825	0.3680	-2.4882	0.0291	0.1331
Christmas$_{10-2}$	0.3357	0.1371	-0.0268	-2.6425	0.0112	0.1384
Christmas$_{10-3}$	0.3359	0.1289	-0.3063	-2.8993	0.0008	0.1436
Christmas$_{10-4}$	0.3307	0.0830	-0.4960	-2.8236	-0.0139	0.1329
Noisy_10	0.3252	-0.2398	-2.5961	-4.2100	-0.1030	0.0282

standard deviation is almost twice as large as the one-word intruder topics. This issue does not happen in the other metrics. This issue does not occur in the other metrics.

Table X. Average metrics performance for 10-words topics and their scaled values. The values are in the format: $avg \pm std(scld)$ - average (avg), standard deviation (std), and scaled ($scld$).

Topic	C_V	C_P	C_{UCI}	C_{UMass}	C_{NPMI}	C_A
$Subj_{10-0}$	0.45±0.08(1.00)	0.58±0.15(1.00)	1.36±0.42(1.00)	-2.24±0.43(0.99)	0.13±0.04(1.00)	0.26±0.07(1.00)
Sub_{10-1}	0.43±0.07(0.80)	0.44±0.15(0.84)	0.96±0.38(0.90)	-2.22±0.24(1.00)	0.09±0.04(0.85)	0.22±0.06(0.83)
$Subj_{10-2}$	0.40±0.05(0.63)	0.29±0.09(0.65)	0.43±0.38(0.76)	-2.38±0.22(0.92)	0.05±0.03(0.68)	0.20±0.05(0.73)
$Subj_{10-3}$	0.39±0.05(0.53)	0.22±0.05(0.56)	0.21±0.36(0.71)	-2.45±0.32(0.89)	0.04±0.02(0.61)	0.18±0.03(0.65)
$Subj_{10-4}$	0.37±0.05(0.36)	0,11±0.03(0.43)	-0.23±0.43(0.60)	-2.68±0.23(0.77)	0.01±0.02(0.48)	0.16±0.02(0.57)
$Noisy_{10}$	0.33±0.00(0.00)	-0,24±0.00(0.00)	-2.60±0.00(0.00)	-4.21±0.00(0.00)	-0.10±0.00(0.00)	0.03±0.00(0.00)

Figure 5 plots the normalized results of the metrics in all 10-word topics. Remark that $Subj_{10-0}$ and $Noisy_{10}$ represent the upper and lower limits of the metrics' performance, respectively. Regarding 5-word topics, C_V and C_P present the same behavior/sensitiveness using 10-words topics. It shows a pattern of both metrics to evaluate topics.

C_{NPMI} presents the same behavior as C_V and C_P, being more similar to C_P. C_A and C_{UCI} show similar behavior regarding 2, 3, and 4 intruder words. The less sensitive metric, again, is C_{UMass}. Despite that, the metric has a low variation among the noisy topics, the worse average result being around 0.74 ($Subj_{10-4}$). Note that the average result of C_V in $Subj_{10-4}$ is 0.33.

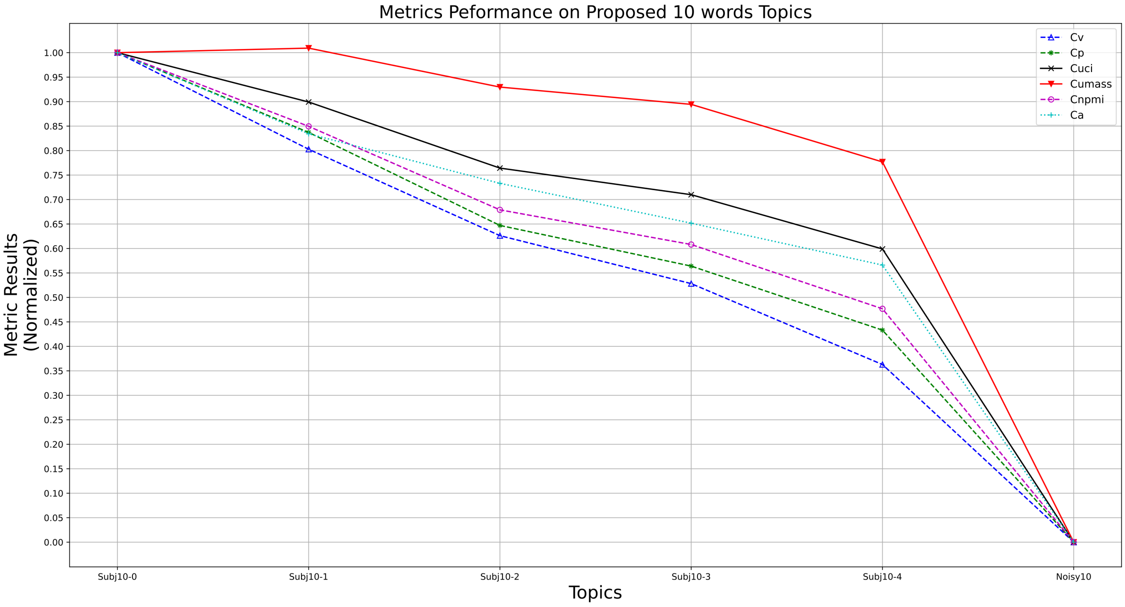


Fig. 5. The average performance of the metrics on the proposed 10 words topics.

5.3 Final Discussion

An overall analysis of the performance of the metrics shows that:

—Although C_{UMass} presents a minor inconsistency against the religion topic, we can claim that the metrics behave on average as expected, *i.e.*, from $Subject_{n-0}$ to $Noisy_n$, the values decrease accordingly.

—If the developers want to compare the discovered topics among different topic model approaches, disregarding "small" noises, C_{UMass} and C_{UCI} are the better choices since the intruder words do not affect the results considering the pure topic as the baseline. C_{UMass} gets better average results on 10-words topics, although its erratic behavior on $Religion_{10-m}$, *i.e.*, all noisy topics get a better result than the pure one; however, the variations are slight (the biggest one in the normalized data is 0.13). See Table IX, column C_{UMass}.

—C_V shows to be the more sensitive metric. In all cases, the results are affected by the number of intruder words. Thus, if the developers want to identify noises on topics, it is the best metric. In the second place, C_P shows to be sensitive as well. C_V and C_P present the most significant difference between pure topics and topics $Subject_{5-2}$ and $Subject_{10-4}$.

—C_A and C_{NPMI} present two different behaviors. On 5-word topics, C_A is more similar to the C_P and C_V, C_{NPMI}, on the other hand, is more similar to C_{UMass} and C_{UCI}. However, on 10-word topics, the behavior is inverted. Therefore, if the developer wants to consider the noises to analyze the discovered topics, those metrics are not the best choices.

6. CONCLUSION

Metrics are essential tools to assess machine learning models. It is a challenging task for unsupervised models since there are no labels to guide the assessment. Another point to consider is how a given metric behaves under certain conditions. In this context, we propose an exploratory analysis to identify how state-of-the-art coherence metrics behave under topics from high to low quality. Based on the metrics performance, we show the behavior under two sets of topics: 5 and 10-word topics. Five subjects were chosen to proceed with the evaluation, and pure and noisy topics were built to accomplish the analysis.

Our analyses demonstrate that the metrics C_V and C_P are more sensitive to noise. That confirms their applicability in scenarios where the user wants to highlight topics with some unrelated words. On the other hand, C_{UMass} and C_{UCI} are more resilient to dirty data and suffer less from noisy information. Such metrics may be used when users want to identify purer topics out of the discovered topics.

Future work intends to extend the analysis by adding more subjects as topics and intruder words and increasing the number of words in the topics, *e.g.*, 15 and 20-words topics. Another direction is to build rare topics (*e.g.*, medical ones) because the intruder words would be more infrequent in a domain-specific subject. We also intend to investigate the metrics' behavior regarding so-called purity metrics (*e.g.*, Jaccard and Gini), besides analyzing the correlation between the proposed topics, their top-n words, and the behavior of coherence metrics.

REFERENCES

Aletras, N. and Stevenson, M. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. pp. 13–22, 2013.

Alvarez, S. A. An exact analytical relation among recall, precision, and classification accuracy in information retrieval. Tech. rep., Boston College, Boston, Technical Report BCCS-02-01, 2002.

Blei, D. M. Probabilistic topic models. *Communications of the ACM* 55 (4): 77–84, 2012.

Bouma, G. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* vol. 30, pp. 31–40, 2009.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. Reading tea leaves: How humans interpret topic models. In *Proceedings of the Twenty-third Advances in neural information processing systems*. pp. 288–296, 2009.

Duarte, D. and Ståhl, N. Machine learning: a concise overview. In *Data Science in Practice*, A. Said and V. Torra (Eds.). Springer, pp. 27–58, 2019.

Fatourechi, M., Ward, R. K., Mason, S. G., Huggins, J., Schlögl, A., and Birch, G. E. Comparison of evaluation metrics in classification applications with imbalanced datasets. In *2008 Seventh International Conference on Machine Learning and Applications*. pp. 777–782, 2008.

Fitelson, B. A probabilistic theory of coherence. *Analysis* 63 (3): 194–199, 2003.

Folleco, A., Khoshgoftaar, T. M., and Napolitano, A. Comparison of four performance metrics for evaluating sampling techniques for low quality class-imbalanced data. In *2008 Seventh International Conference on Machine Learning and Applications*. pp. 153–158, 2008.

Juba, B. and Le, H. S. Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. International Joint Conferences on Artificial Intelligence, pp. 4039–4048, 2019.

Lau, J. H., Newman, D., and Baldwin, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. pp. 530–539, 2014.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology* 22 (140): 65–68, 1932.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 262–272, 2011.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, USA, pp. 100–108, 2010.

Nikolenko, S. I. Topic quality metrics based on distributed word representations. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*. pp. 1029–1032, 2016.

O'Callaghan, D., Greene, D., Carthy, J., and Cunningham, P. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications* 42 (13): 5645–5657, 2015.

Powers, D. M. W. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2 (1): 37, 2011.

Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, USA, pp. 399–408, 2015.

Steyvers, M. and Griffiths, T. Probabilistic topic models. In *Handbook of latent semantic analysis*, T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Eds.). Laurence Erlbaum Associates, 21, pp. 424–440, 2007.

Vijaymeena, M. and Kavitha, K. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal* 3 (2): 19–28, 2016.