# Atividade 05 - Resultados Finais

## Equipe

**Nome:** Detetives Virtuais
**Repositório GitLab da equipe:** https://gitlab.com/jr_waine/csb-51
**Membros:**

- Waine Barbosa de Oliveira Junior, 1905120,
  jr_waine, waine@alunos.utfpr.edu.br, Eng. Comp, UTFPR
- Thiago Schinda Bubniak, 1540778,
  thiagobubniak, thiagobubniak@alunos.utfpr.edu.br, Eng. Comp, UTFPR

## Tema

Compreender quais os temas de fake news e notícias verdadeiras mais relevantes em determinado período e comparar com eventos relacionados no período.

## Perguntas de pesquisa

- Quais foram os tópicos mais comentados nas notícias?
- Como esses tópicos evoluíram com o passar do tempo?
- Qual a relação desses tópicos com eventos políticos?

## Hipóteses

- Os tópicos das notícias evoluíram conforme eventos políticos (e.g. perto do julgamento do ex-presidente, as notícias falsas sobre eles aumentaram)
- Os tópicos das notícias falsas e verdadeiras, num mesmo intervalo de tempo, tendem a ser os mesmos

## Dados

Com relação aos dados, utilizamos o corpus Fake.br como base. O corpus possui 7200 notícias retiradas de sites, sendo 50% delas verdadeiras e 50% falsas. As categorias das notícias são apresentadas a seguir.

| | |
|---|---|
| politica | 4180 |
| tv_celebridades | 1544 |
| sociedade_cotidiano | 1276 |
| ciencia_tecnologia | 112 |
| economia | 44 |
| religiao | 44 |

É possível perceber uma concentração em notícias sobre política, com essas sendo mais de 50% do corpus.

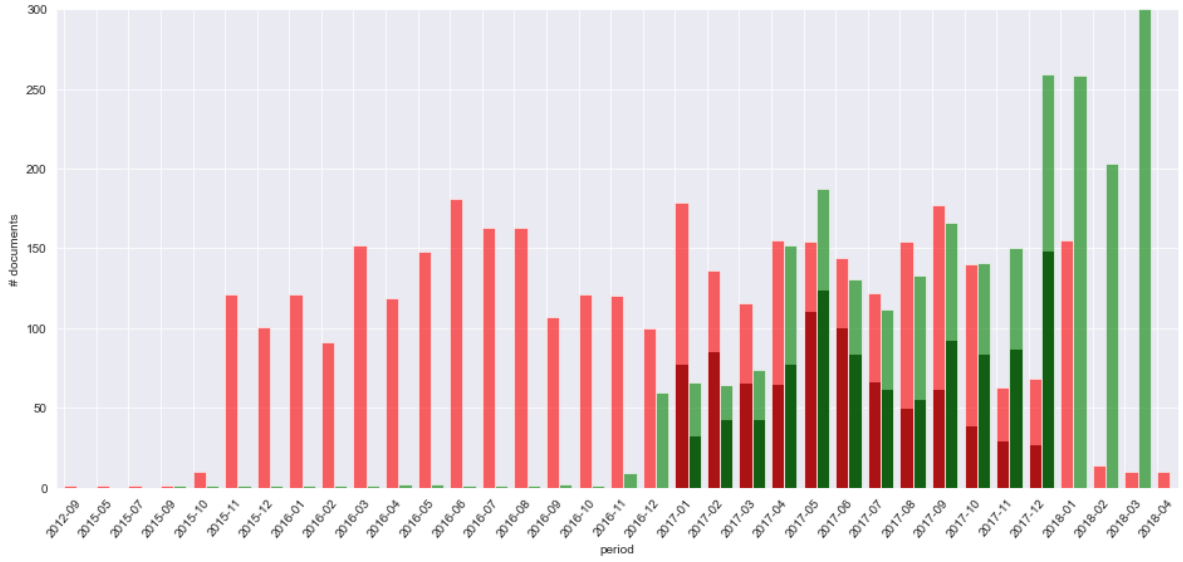
Com relação aos sites fontes das notícias do corpus, nenhum site teve disseminação de ambas as notícias (verdadeiras e falsas), ou um, ou outro. É possível perceber que há uma concentração de sites de teor político, o que, tendo em vista o proposto, faz com que haja uma boa base de dados para análise.

| link | label | |
|---|---|---|
| diariodobrasil.org | fake | 3337 |
| g1.globo.com | true | 2300 |
| politica.estadao.com.br | true | 753 |
| afolhabrasil.com.br | fake | 174 |
| internacional.estadao.com.br | true | 114 |
| cultura.estadao.com.br | true | 95 |
| www1.folha.uol.com.br | true | 91 |
| thejornalbrasil.com.br | fake | 66 |
| economia.estadao.com.br | true | 51 |
| brasil.estadao.com.br | true | 46 |
| esportes.estadao.com.br | true | 30 |
| alias.estadao.com.br | true | 29 |
| ceticismopolitico.com | fake | 16 |
| sao-paulo.estadao.com.br | true | 16 |
| saude.estadao.com.br | true | 15 |
| sustentabilidade.estadao.com.br | true | 13 |
| opiniao.estadao.com.br | true | 11 |
| topfivetv.com | fake | 7 |
| estadao.com.br | true | 7 |
| educacao.estadao.com.br | true | 7 |
| ciencia.estadao.com.br | true | 5 |
| viagem.estadao.com.br | true | 5 |
| emais.estadao.com.br | true | 3 |
| link.estadao.com.br | true | 2 |
| f5.folha.uol.com.br | true | 2 |
| territorioeldorado.limao.com.br | true | 2 |
| datafolha.folha.uol.com.br | true | 1 |
| blogdofred.blogfolha.uol.com.br | true | 1 |
| acervo.estadao.com.br | true | 1 |

dtype: int64

Sites fontes das notícias do corpus

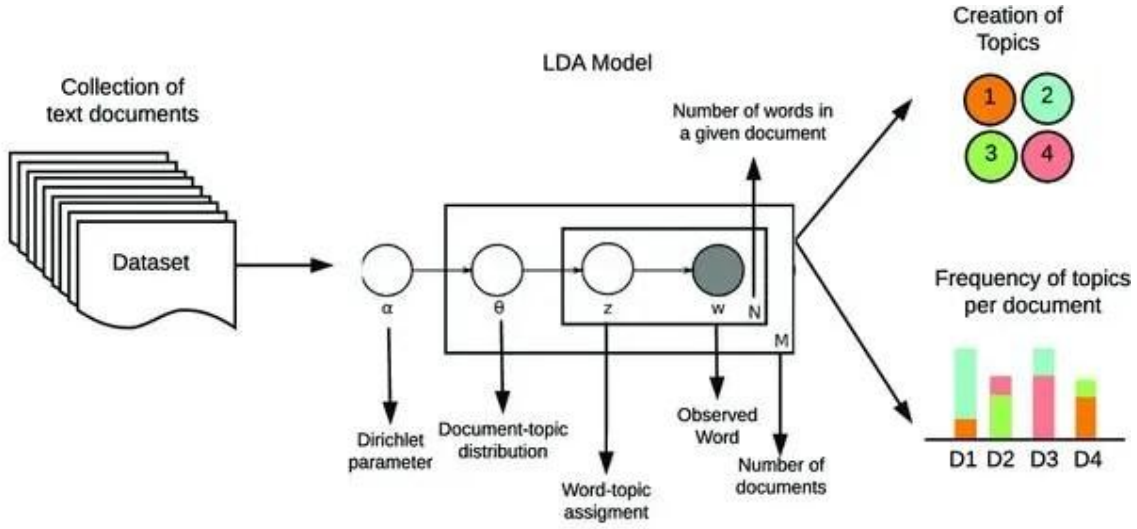
Também foi feita uma análise temporal das notícias no corpus. Nota-se que as notícias falsas são distribuídas de maneira quase uniforme pelo período do final de 2015 até o início de 2018. Enquanto isso, as notícias verdadeiras se concentram do início de 2017 até o início de 2018. Para limitar o escopo dos tópicos das notícias, ao mesmo tempo mantendo uma grande quantidade de dados, restringimos nossa análise ao ano de 2017.




Além dessa restrição temporal, também limitamos a categoria das notícias utilizadas para aquelas de política. Sendo assim, foram utilizados um total de 1718 notícias (~24% do corpus original) nas nossas análises, sendo 936 verdadeiras e 782 falsas.

# Modelos

Para a detecção de tópicos, utilizamos o *LDA (latent Dirichlet allocation) (Blei et. al)*. O modelo considera como entrada uma série de documentos e, a partir disso, gera tópicos a partir de tais documentos (quantidade é definida previamente), com cada documento tendo uma certa frequência de cada tópico, sendo cada um representado por um conjunto de palavras.




Representação do LDA (*Latent Dirichlet Allocation*)

O modelo não considera a dimensão temporal dos documentos, é importante manter isso em mente para a análise dos tópicos ao longo do tempo. Tendo em vista que os documentos (notícias, no nosso caso) foram limitados ao ano de 2017, julgamos que os tópicos presentes ocorreram durante o ano todo.

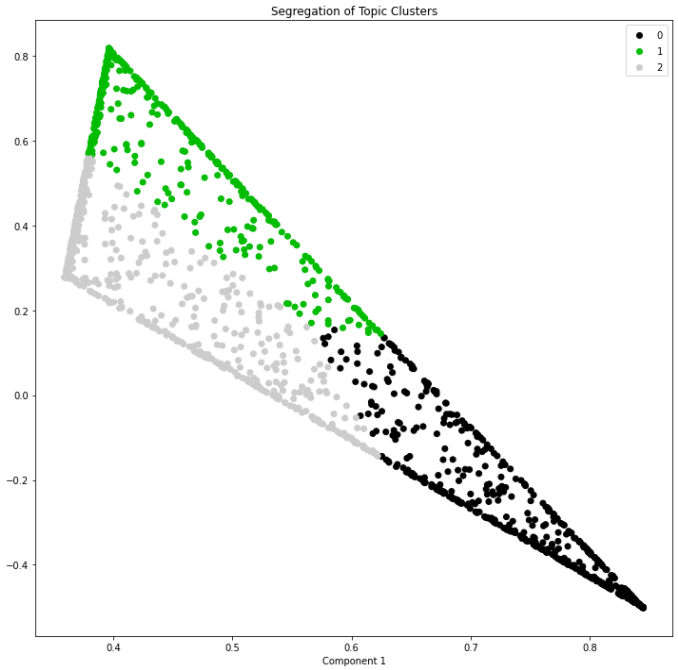
Para a preparação dos dados, foram utilizados diversos métodos de tratamento de texto a fim de facilitar e melhorar o desempenho das análises e da aplicação de modelos como o LDA. Em resumo, os pré processamento aplicados foram:

- Remoção de acentos e pontuação;
- Remoção de stopwords;
- Normalização de texto usando NFKD;
- Tokenização e lematização;
- Criação de matriz termo-documento.

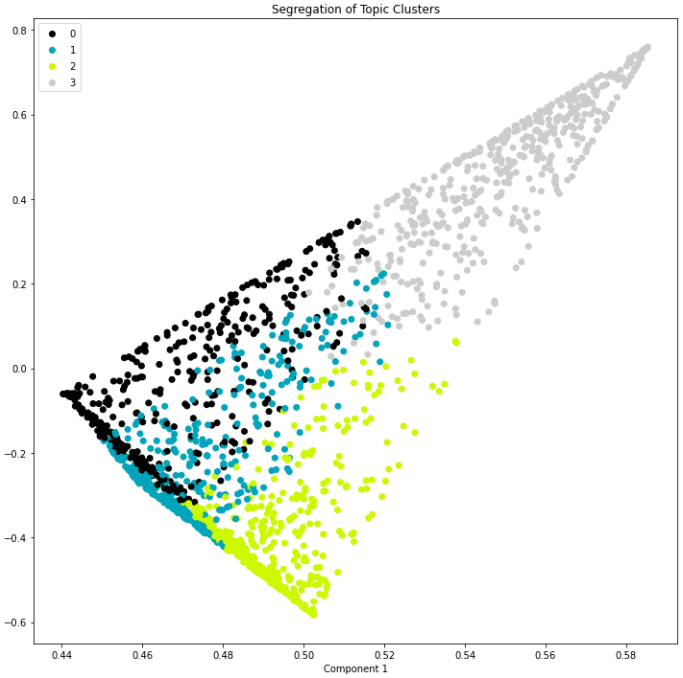
É importante ressaltar também que, no caso da lematização, apesar de bastante útil para reduzir variações de uma mesma palavra em uma única palavra comum, pôde-se perceber a redução errada em alguns casos, tal como “Moro”, sobrenome do Juiz Sérgio Moro, sendo reduzido para “morar”.

# Resultados

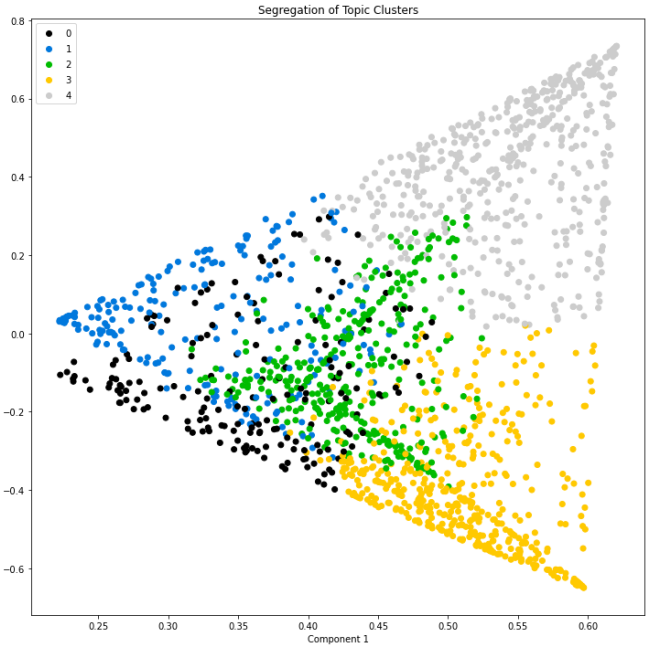
A utilização do método LDA para a descoberta de tópicos foi realizada com a ajuda do GridSearch, ferramenta usada para ajudar a definir a melhor quantidade de tópicos e parâmetros para o LDA, que, juntamente com a técnica de SVD (*Singular Value Decomposition*) a qual fornece uma avaliação visual da distribuição dos documentos pelos grupos, foram obtidos os seguintes resultados:




LDA - 3 Tópicos. Log Likelihood: -621096.




LDA - 4 Tópicos. Log Likelihood: -624739.

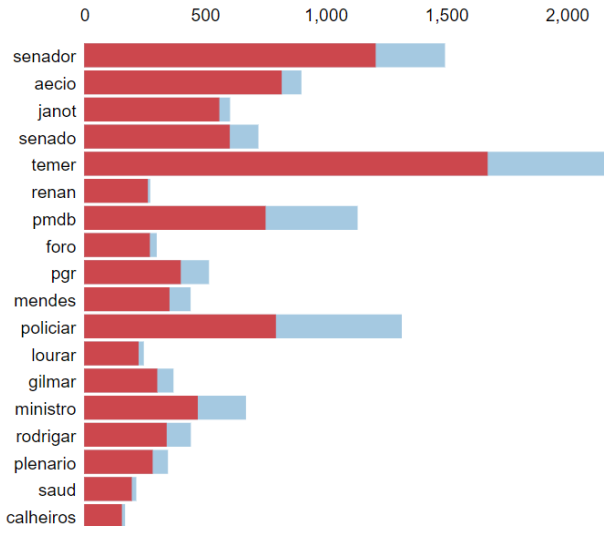



LDA - 5 Tópicos. Log Likelihood: 628583.

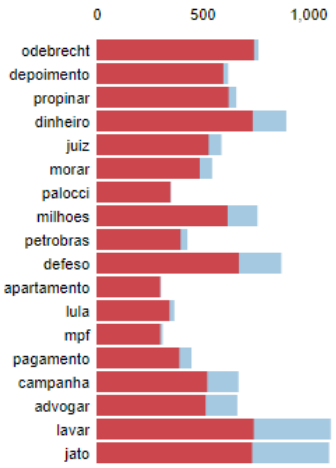
Tendo em vista o resultado numérico e também o qualitativo, decidimos manter 3 tópicos apenas. Abaixo são apresentadas as características gerais dos tópicos.

| | Word 0 | Word 1 | Word 2 | Word 3 | Word 4 | Word 5 | Word 6 | Word 7 | Word 8 | Word 9 | Word 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Topic 0** | temer | senador | presidente | ministrar | aecio | federal | policiar | deputar | pmdb | denunciar | pedir |
| **Topic 1** | presidente | federal | pedir | afirmar | odebrecht | lavar | dinheiro | jato | defeso | empresar | propinar |
| **Topic 2** | presidente | governar | pai | temer | politicar | deputar | pessoa | afirmar | publicar | brasileiro | passar |

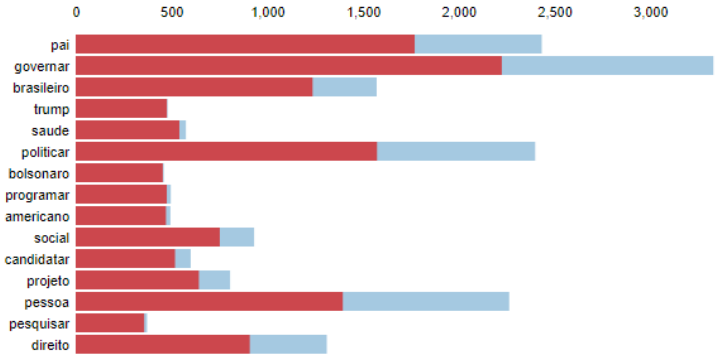
Palavras mais comuns por tópico




Palavras representativas do tópico 0




Palavras representativas do tópico 1




Palavras representativas do tópico 2

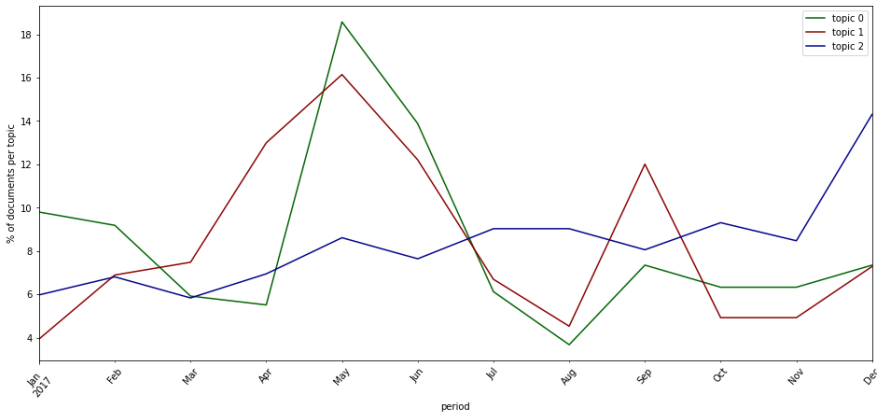
| dominant_topic | label | |
|---|---|---|
| 0 | fake | 252 |
| | true | 238 |
| 1 | fake | 193 |
| | true | 315 |
| 2 | fake | 337 |
| | true | 383 |

Número de notícias por tópico (considerando tópico dominante)

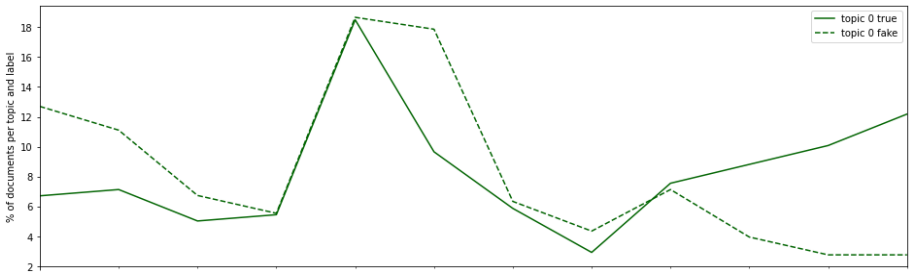
Levando em conta as palavras mais representativas e também o intervalo de tempos, definimos em um âmbito maior os tópicos como:

- **Tópico 0:** Notícias relacionadas ao presidente Temer e a deputados e senadores
- **Tópico 1:** Notícias sobre a investigação Lava-Jato
- **Tópico 2:** Notícias sobre os EUA e pré-candidatura de Bolsonaro (discurso parecido com o de Trump)

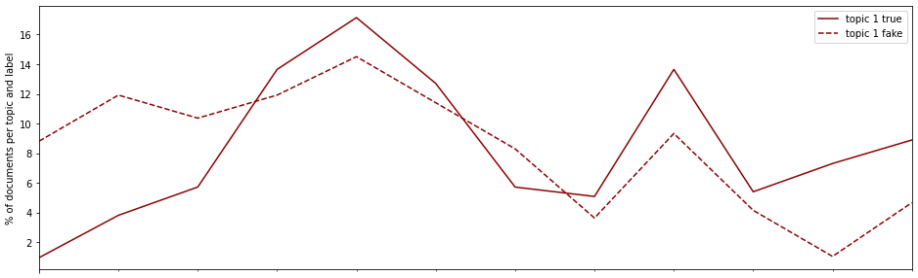
Outro aspecto importante dos tópicos é sua frequência ao longo do ano de 2017. As imagens a seguir mostram essa análise.



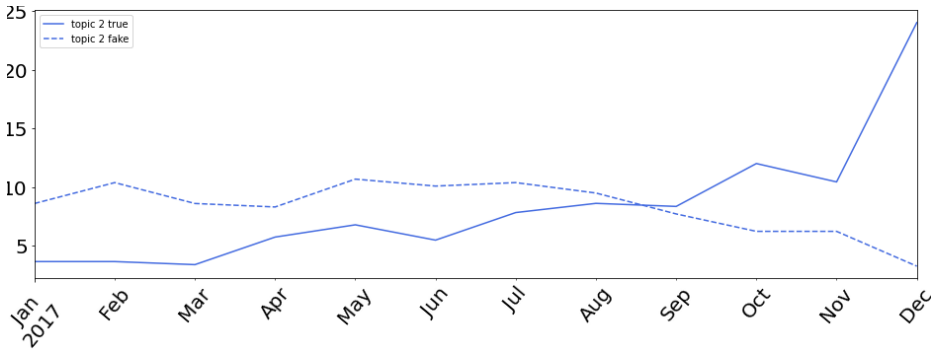

Frequência normalizada ao longo de 2017 dos tópicos




Frequência normalizada do tópico 0




Frequência normalizada do tópico 1




Frequência normalizada do tópico 2

No tópico 0, que consiste em notícias sobre a presidência e Câmara/Senado, os picos e os eventos encontrados foram:

- Janeiro: Morte de Teori Zavaski
- Maio: divulgação do áudio de Joesley Batista com Michel Temer

- Setembro: Afastamento de Aecio Neves

No tópico 1, com notícias sobre a Lava-Jato, os picos e os eventos encontrados foram:

- Abril: Projeto de abuso de autoridade é aprovado
- Julho: Lula condenado a 9 anos por Moro
- Setembro: 51 milhões de Geddel

No tópico 2, com notícias sobre Trump e Bolsonaro, os picos e os eventos encontrados foram:

- Janeiro: Trump assume a presidência dos EUA
- Dezembro: Trump reconhece Jerusalém como capital de Israel
- Ao longo do ano: Trump e Bolsonaro quebrando protocolos, comparações entre os dois.

Apesar das relações apresentadas entre tópicos e fatos, há vários fatos que ocorreram no período e não causaram mudanças significativas nos tópicos. As várias reformas feitas em 2017 (trabalhista, previdência, política) são exemplos.
Na análise foi feito o caminho reverso de, a partir da mudança da frequência de tópicos, encontrar uma justificativa para tal. Acreditamos que encontrar a tendência de tópicos a partir de eventos políticos seja tão ou mais frutífero que o processo que foi feito.

Com relação às tendências de notícias verdadeiras e falsas, em alguns casos são semelhantes, já em outros não. Não é possível fazer uma conclusão forte sobre a relação entre notícias verdadeiras/falsas e os tópicos, pois isso pode ser uma coincidência das notícias obtidas pelo corpus. Para uma análise mais detalhada, seria necessário a definição de uma metodologia com relação à obtenção de notícias e seus tópicos.

# Limitações

A análise a partir do LDA não engloba o escopo temporal. Um modelo como *Topics Over Time (Wang, McCallum)* para a evolução de tópicos no tempo deve permitir uma melhor análise para intervalos de tempos maiores.
Também, tendo em vista o interesse em tópicos e tendências, utilizar como fonte tweets ou outras fontes mais “instantâneas” deve refletir melhor a volatilidade dos tópicos com o passar do tempo.
Com relação ao corpus utilizado, não é apresentada uma metodologia para obtenção dos dados, o que pode fazer com que os resultados obtidos sejam incidentais.
Por fim, a relação entre tópicos e eventos foi feita na mão, tendo em vista o pequeno número de tópicos e também o intervalo relativamente curto. Porém um modelo para relacionar tópicos e eventos políticos seria uma grande contribuição para análises nesse sentido, permitindo assim analisar um maior número de tópicos, assim como um maior intervalo de tempo.