Construção de Datasets para Segmentação Automática de Hashtags

Juliana Resplande, Ruan Rodrigues, Marcelo Inuzuka, Acquila Rocha, Hugo do Nascimento

Resumo Nesse artigo, analisou-se um procedimento simples para construção de *datasets* de segmentação de *hashtags* a partir dos microtextos de *dumps* de *tweets*. Foram geradas segmentações automáticas por meio de um algoritmo heurístico, no qual verificamos estatisticamente que esse método consegue corrigir anomalias de modelos avançados de segmentação, podendo melhorar as novas arquiteturas de *deep learning* nessa tarefa.

1 Introdução

A *segmentação de hashtags* é um passo frequentemente utilizado na tarefa de pré-processamento em *pipelines* em conjuntos de dados de redes sociais, no sentido de melhorar o desempenho de análise de sentimentos e a detecção de eventos. É também um caso específico da segmentação de palavras, cujo objetivo é a inserção de caracteres separadores entre as palavras de uma sentença quando estes foram suprimidos por algum motivo anormal, tais como erros na conversão de imagens em texto.

Para um aprendizado de máquina efetivo, é fundamental ter disponível um conjuntos de dados de boa qualidade, pois isso afeta diretamente nos processos de treinamento de avaliação de modelos. Especificamente para a tarefa de segmentação de *hashtags*, foi constatado pelos autores vários problemas nos *datasets* atuais, tais como: escassez, baixo volume de exemplos, falta de avaliação de qualidade de anotação, falta de informação temporal e baixa diversidade de línguas.

Além destes problemas, há um desafio inerente ao contexto de aplicação em redes sociais: a alta presença de palavras fora do vocabulário, como gírias e nomes próprios, o que acelera a

Juliana Resplande
Instituto de Informática – Universidade Federal de Goiás (UFG), Goiânia, Goiás, Brasil.
e-mail: julianarsg13@gmail.com

Ruan Rodrigues
Instituto de Informática – Universidade Federal de Goiás (UFG), Goiânia, Goiás, Brasil.
e-mail: ruanchaves93@gmail.com

Marcelo Inuzuka
Instituto de Informática – Universidade Federal de Goiás (UFG), Goiânia, Goiás, Brasil.
e-mail: marceloakira@ufg.br

Acquila Rocha
Instituto de Informática – Universidade Federal de Goiás (UFG), Goiânia, Goiás, Brasil.
e-mail: acquila@discente.ufg.br

Hugo do Nascimento
Instituto de Informática – Universidade Federal de Goiás (UFG), Goiânia, Goiás, Brasil.
e-mail: hadn@inf.ufg.br

Anais do XV Encontro Anual de Ciência da Computação (EnAComp 2020). ISSN: 2178-6992.
Catalão, Goiás, Brasil. 25 a 27 de Novembro de 2020.

depreciação dos *datasets* e também leva a dúvidas sobre o desempenho de generalização dos modelos em contexto temporais distintos. Tal desafio é exemplificado na Tabela 1. Com muita frequência, as hashtags contém entidades nomeadas cuja relevância na rede social está altamente condicionada ao seu período de maior popularidade: seja ela uma banda, uma convenção, uma tecnologia, e assim por diante.

Hashtag	Segmentação correta	Significado
#BTSARMY	BTS ARMY	Fã clube da banda coreana BTS.
#KCONJAPAN	KCON JAPAN	Convenção anual de K-pop realizada no Japão.
#VMwareNSX	VMware NSX	Produto da empresa de tecnologia VMware.

Tabela 1: Exemplos de hashtags comuns no dataset construído.

Considerando-se estes problemas e desafios, foi proposto neste trabalho um método automático e otimizado para a construção de *datasets* de segmentação de *hashtags*, extraído de *dumps* de mensagens do Twitter publicamente disponíveis. Para avaliar a efetividade do método, realizaram-se testes de concordância de anotação utilizando a métrica de Cohen's Kappa (BANERJEE et al., 1999). Os resultados do método foram confrontados por meio de um segmentador automatizado, GPT-2 (RADFORD et al., 2019), no qual 1666 instâncias houveram discordância entre o segmentador e a heurística. Esses exemplos destoantes foram adicionalmente avaliados por dois anotadores humanos. Como resultado, foi obtido a segmentação de 156.170 hashtags únicas na língua inglesa, extraídas de 2.463.116 *tweets*.

Este trabalho tem como objetivo avaliar e aperfeiçoar o método automático de construção de um conjunto de dados destinados a tarefa de segmentação de *hashtags*, proposto inicialmente por (ÇELEBI; ÖZGÜR, 2016), bem como produzir um *dataset* de boa qualidade e livremente disponível para a comunidade científica.

2 Trabalhos Relacionados

Çelebi e Özgür (2016) considerou dois métodos para construção de três conjuntos de dados destinados à segmentação de *hashtags*: (D1) SNAP (803K exemplos), (D2) Stadford (1000 exemplos) e (D3) BOUN (1000 exemplos). O primeiro método originou D1 e envolveu extrair automaticamente milhares de *hashtags* do conjunto de dados SNAP Stanford Twitter (YANG; LESKOVEC, 2011), nas quais foram aplicadas heurísticas de segmentação automática de palavras. O funcionamento das heurísticas consistiu em buscar, no conjunto de dados SNAP, palavras mais frequentes que compunham a *hashtag* pesquisada. O segundo método originou D2 e D3 e envolveu a construção de um conjunto de dados artificial, no qual cada *hashtag* podia ser constituída da junção pela palavras extraídas dos *tweets* em duas fontes distintas: Stanford Sentiment Data e Twitter Search API.

Por outro lado, Maddela, Xu e Preoţiuc-Pietro (2019) propuseram um conjunto de dados manualmente anotado composto por 12.594 *hashtags*, bem como uma abordagem de segmentação de *hashtags* que encara a segmentação como um problema de classificação de pares. A construção do conjunto de dados envolveu uma série de anotadores humanos, sujeitos à um processo de

controle de qualidade utilizando *hashtags* de teste. Desta forma, os anotadores eram desclassificados caso errassem mais de 20% dos exemplos de teste.

A literatura apresentada aborda a construção de conjuntos de dados para segmentação de *hashtags*, com a utilização de dois procedimentos primordiais de anotação: manual e automático. No entanto, até o conhecimento dos autores, esta pesquisa é a primeira em avaliar as divergências entre anotadores humanos e as abordagens computacionais por concordância das anotações.

3 Metologia

O *pipeline* dos conjuntos de dados baseou-se em Çelebi e Özgür (2016). As *hashtags* e os *tweets* foram extraídos de períodos distintos e verificou-se possíveis segmentações das *hashtags* nos *tweets*. O processo operou da seguinte forma: inicialmente são escolhidos dias para baixar os *dumps*; a partir dos arquivos as informações relevantes são extraídas gerando uma tabela de dados de ***tweets*** e outra de ***hashtags***; por fim, a segmentação heurística é executada em cima das tabulações obtidas. O processo é ilustrado na figura 1 abaixo:

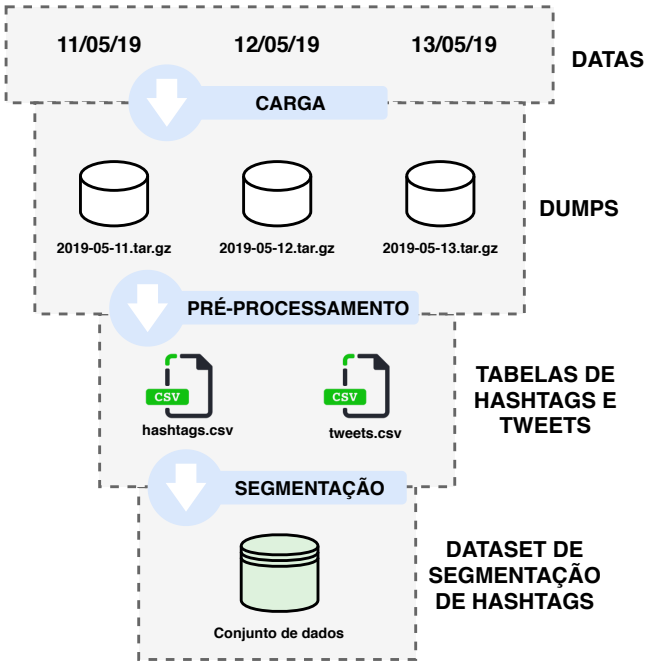

Figura 1: *Pipeline* em um período de 3 dias.

Foram utilizados 2,46 milhões de *tweets* e segmentados 15 mil *hashtags* em inglês neles presentes. As rotinas de processamento foram desenvolvidos inicialmente com o *Google Colabolatory*, que disponibiliza gratuitamente máquinas com apenas uma CPU para os experimentos. A fim de atender uma necessidade de paralelismo durante a segmentação heurística, utilizou-se posteriormente uma máquina com nó de 80 núcleos com o tempo de computação limitado a 12 horas.

3.1 Carga

Os *dumps* diários do Twitter foram coletados do serviço de arquivamento fornecido pela equipe Twitter Stream [1]. Os períodos escolhidos para a carga foram 01-07/11/2018, 11-17/05/2019 e 25-31/01/2020, sendo a escolha das datas arbitrária. Cada dia resultou em um arquivo comprimido `tar.gz` com cerca de 1,9 GB contendo documentos `json` comprimidos no formato `bz2` sobre *tweets* para cada segundo.

Essa etapa de carga representou o maior gargalo no *pipeline*. O tempo de *download* de cada *dump* foi de aproximadamente 1 hora. Ou seja, para baixar os 21 dias mencionados acima, levou-se 21 horas, mesmo utilizando-se uma rede acadêmica de alta velocidade. Para solucionar esse problema será testado futuramente o *download* empregando um conjunto de *proxies* ou a utilização protocolo *torrent*.

3.2 Pré-processamento

A partir dos múltiplos arquivos `json` baixados, foram extraídos os campos relevantes, como os textos dos *tweets* (`'text'`) e as *hashtags* (`['entities']['hashtags']`) e como o idioma dos microtextos (`'lang'`). Informações sobre os campos fornecidos pelo `json` estão disponíveis na página do Twitter [2].

Os microtextos, posteriormente, foram filtrados em inglês, francês, português ou espanhol. No *dump* de 14/05/2019, cerca de 34,65% dos *tweets* estavam em algum desses idiomas. Foram selecionados também *hashtags* desses trechos que utilizassem somente com letras do alfabeto latino. Por volta de 4,08% dos microtextos na data mencionada, possuíam *hashtags* que satisfaziam essa condição.

Em seguida, foram geradas dois arquivos `csv`, um com os *tweets* e o outro com as *hashtags*. A tabela dos *tweets* possui campos com o texto do *tweet* e o seu idioma, para ser utilizada na busca da segmentação heurística. A tabela de *hashtags* é mais completa e contém as colunas de texto da *hashtag*, id do *tweet* da *hashtag*, texto do *tweet*, idioma do *tweet* e o *timestamp* do *tweet* em segundos. As tabelas de *hashtags* e *tweets* gerados de um *dump* diário ocupam respectivamente cerca de 26MB e 48MB.

As tabelas sobre *tweets* e *hashtags* diárias são concatenadas formando duas tabelas de *hashtags* e *tweets* de um certo período.

3.3 Segmentação heurística

O último passo da *pipeline* é a segmentação heurística a partir desses dois documentos. Procura-se possíveis segmentações para cada *hashtag* ao longo de todos os *tweets*. Com o objetivo de facilitar o processamento em paralelo, os dados nos três períodos foram subdivididos em seis janelas menores de três dias casa. A saber: 01-03/11/2018, 04-06/11/2018, 11-13/05/2019, 14-16/05/2019, 25-27/01/2020, 27-30/01/2020. Além disso, reduzimos a línguas de segmentação

[1] https://archive.org/details/twitterstream

[2] https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/intro-to-tweet-json

inicialmente de inglês, francês, espanhol e português para somente em inglês, uma vez a língua inglesa compreende cerca de 76% das *hashtags* enquanto a língua portuguesa possui 4% das *hashtags*.

Algoritmo 2: Segmentação heurística de uma *hashtag*.

Entrada: hashtag h, lista T de *tweets*
Saída:

```
1  Seja R uma expressão regular vazia.
2  para cada letra l em h faça
3      se l não for a letra última de h então
4          R ← R + l + ␣*
5      senão
6          R ← R + l
7      fim
8  fim
9  Pré-carregue o autômato de R de forma uncased.
10 Seja L uma lista vazia.
11 para cada tweet t em T faça
12     L ← L + todas cadeias em t que satisfaçam R
13 fim
14 retorna termo mais frequente em L ou NULO se L = ∅
```

A segmentação é realizada por meio de expressão regular. O algoritmo 2 acima indica como foi executado a segmentação heurística. O espaço é retratado por ␣, enquanto a concatenação é representada pelo símbolo de soma (+). O algoritmo é significativamente otimizado ao pré-carregar o autômato de R antes da utilização dessa expressão regular.

A figura 2 ilustra o algoritmo 2 da segmentação heurística. A entrada '#GameOfThrones', gera uma expressão regular `#␣*g␣*a␣*m␣*e␣*o␣*f␣*t␣*h␣*r␣*o␣*n␣*e␣*s`, cujo autômato é pre-carregado de forma *uncased*, isto é, ignorando maiúsculos e minúsculos. Posteriormente, procura-se em todos tweets palavras que satisfaçam o autômato gerado. Todos os *matchings* são contados a fim de encontrar a segmentação mais frequente, `game of thrones`, a qual é retornada.

Para mais, o algoritmo foi executado, como um todo, de forma paralela em vários núcleos. Mesmo não mencionado pelo Çelebi e Özgür (2016), optou-se por fazer a segmentação *uncased*, porque em testes iniciais com a língua portuguesa ignorar maiúsculos e minúsculos aumentou em 10% o número de *hashtags* com sugestões de segmentação.

4 Outros métodos de segmentação

Para verificar a qualidade da segmentação das *hashtags*, outras abordagens complementares foram utilizadas: um método baseado no modelo GPT-2 e a segmentação humana.

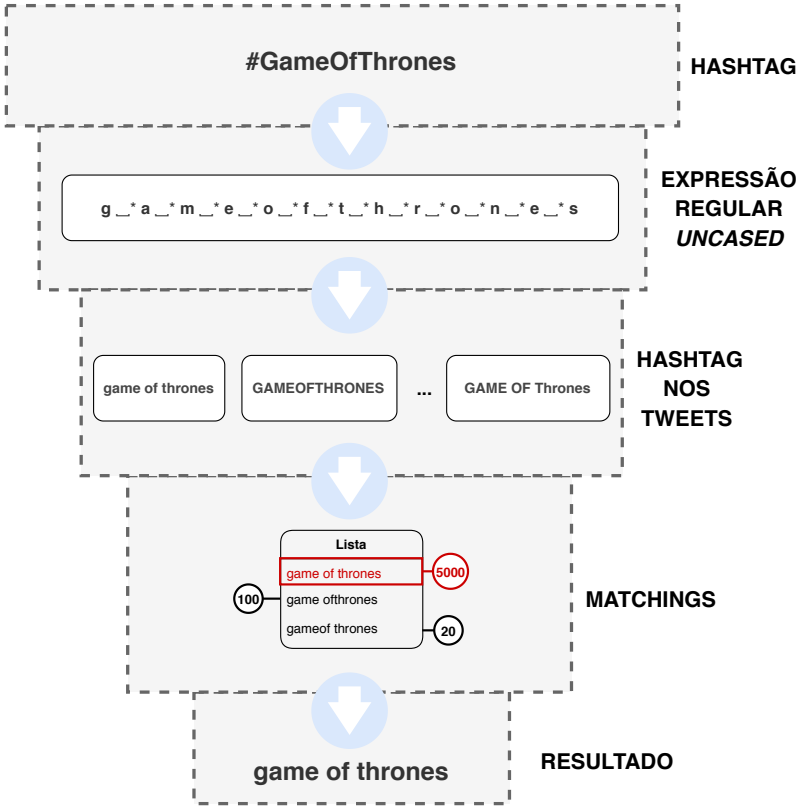

Figura 2: Exemplo de execução para a hashtag '#GameOfThrones'

4.1 Segmentação automática com GPT-2

GPT-2 é um modelo de linguagem (RADFORD et al., 2019) que adota a arquitetura de *Transformers* e faz o uso exclusivamente de mecanismos de atenção (VASWANI et al., 2017). Esta arquitetura permite a aplicação de *transfer learning*: isto é, que um modelo pré-treinado em uma extensa base textual seja imediatamente aplicado em novos contextos.

Embora seja possível realizar um ajuste fino (*fine-tuning*) do modelo GPT-2 à tarefa de segmentação de *hashtags*, optou-se por aplicar o modelo à tarefa sem fornecê-lo qualquer conhecimento prévio além daquele já acumulado na etapa de pré-treino. Foi utilizada a versão mais compacta do modelo, o GPT-2 *small*, que possui 117 bilhões de parâmetros.

Seguindo trabalhos anteriores na área de segmentação de palavras com modelos de linguagem, adotamos a abordagem *Beamsearch* (DOVAL; GÓMEZ-RODRÍGUEZ, 2018), a qual consiste em encontrar a melhor segmentação possível a partir de uma busca no espaço de possíveis soluções, onde as diversas possibilidades são ranqueadas por *scores* fornecidos pelo modelo de linguagem. Embora tais trabalhos anteriores tenham se baseado em redes neurais recorrentes, optamos por substituí-las pelo modelo GPT-2, que apresenta resultados mais próximos ao estado-da-arte em tarefas de compreensão de linguagem natural, que são amplamente documentados na literatura (RADFORD et al., 2019).

4.2 Segmentação humana

A anotação foi realizada por dois indivíduos. O primeiro anotador tinha permissão para consultar sobre a segmentação dos termos na Internet, enquanto o segundo avaliador não podia verificar se separou corretamente os termos da *hashtag*. Desta forma, a análise do primeiro anotador é considerada o *gold standard*.

Ambos avaliadores tinham ciência do domínio dos termos, ou seja, sabiam que iriam segmentar *hashtags* do *Twitter*. Uma regra já definida era que, caso fosse encontrado um termo que normalmente possui apóstrofo, as palavras seriam segmentadas. Assim a *hashtag* ”daddys”, por exemplo, seria separada como ”daddy s”.

Nota-se, contudo, que a segmentação de algumas *hashtags* dependia do contexto. A *hashtag* ”waterdown”poderá ser separada como ”Water down”, o que significaria ”enfraquecer”, ou não será segmentada, quando se referir à cidade Waterdown no Canadá. Ademais, o anotador *gold* observou, ao pesquisar na Internet, que havia uma quantia significativa de *hashtags* que eram entidades nomeadas de lugares, gírias e artistas.

5 Métrica de concordância

Para comparar a divergência entre os anotadores humanos e as abordagens computacionais, avaliamos a concordância das anotações utilizando o coeficiente Kappa (BANERJEE et al., 1999). Dado que o coeficiente kappa compara divergências entre classes apontadas por anotadores, foi necessário formular o problema de segmentação de *hashtags* como um problema de classificação, em particular, como um problema de classificação de caracteres. Isto é, dada a sequência de caracteres de uma *hashtag*, segmentá-la corretamente é equivalente a atribuir, a cada caractere, uma classe que corresponde à sua posição na palavra.

Partindo de um trabalho popular na literatura de segmentação de palavras (ZHAO et al., 2019), consideramos a abordagem BMES. Para essa finalidade, ela consiste em atribuir, a cada caractere de uma hashtag, uma dentre as seguintes classes:

- a classe **B**, caso o caractere ocorra no início (***b**eginning*) de uma palavra com mais de um caractere;
- a classe **M**, caso o caractere ocorra no meio (***m**iddle*) de uma palavra;
- a classe **E**, caso o caractere ocorra ao final (***e**nd*) de uma palavra;
- a classe **S**, caso o caractere seja a única (***s**ingle*) letra da palavra.

A título de exemplo, para uma *hashtag* como *#thexfactor*, podemos tomar a segmentação *”the x factor”*: ela será convertida no trecho da lista com os rótulos $[B,M,E,S,B,M,M,M,M,E]$, conforme demonstrado na Tabela 2. Como resultado final, as segmentações de cada anotador são representadas como uma lista de rótulos de classes posicionais. O coeficiente Kappa, portanto, é calculado a partir da comparação entre estas listas.

t	h	e	x	f	a	c	t	o	r
B	M	E	S	B	M	M	M	M	E

Tabela 2: Rotulação da segmentação *"the x factor"* para a hashtag *#thexfactor* de acordo com o esquema de rotulação de caracteres BMES.

6 Resultados

Nessa seção, são comparadas as segmentações apresentadas pelos anotadores humanos pela abordagem usando modelo GPT-2 e pela técnica de segmentação heurística, empregando o coeficiente Kappa, mencionado anteriormente. Também comparamos em maior profundidade os resultados atingidos pela abordagem com o modelo GPT-2 e a segmentação heurística.

6.1 Comparação pelo coeficiente Kappa

Conforme observamos na Tabela 3, o coeficiente Kappa mais elevado foi atingido entre o anotador humano que realizou consultas (*Gold*) e o anotador humano que realizou anotações espontâneas referenciado aqui apenas por "*Humano*", chegando a um valor muito próximo ao valor máximo, que é 1.0.

	GPT-2	Gold	Heurística
Gold	0.898		
Heurística	0.601	0.679	
Humano	0.897	0.969	0.682

Tabela 3: Coeficientes Kappa calculados dois a dois entre o anotador humano que realizou consultas (*Gold*), o anotador humano espontâneo (*Humano*), a abordagem com o modelo GPT-2 (*GPT-2*) e a segmentação heurística (*Heurística*).

Observamos que a abordagem GPT-2 apresenta um coeficiente Kappa relativamente próximo a ambos anotadores humanos (*Humano* e *Gold*), registrando valores próximos de 0.9.

A segmentação heurística, entretanto, apresenta um coeficiente Kappa em comparação com os anotadores humanos bastante aquém daquilo que é alcançado pela abordagem GPT-2: seus valores oscilam em torno de 0.68.

O modelo GPT-2, por tanto, foi capaz de atingir uma qualidade de segmentação de *hashtags* bastante próxima à humana. Ademais, também observa-se que, dado o volume de dados considerado, a segmentação heurística apresentou uma qualidade significativamente inferior.

6.2 Comparação adicional entre GPT-2 e Heurística

Embora o coeficiente kappa da segmentação heurística mostre um desempenho insatisfatório da técnica tomada isoladamente, a sua comparação com os resultados da abordagem com o GPT-2 por meio de outra análise revelam aspectos bastante pertinentes.

A Figura 3 contém cinco gráficos. Eles são separados por um nível de discordância entre as soluções da GPT-2 e da segmentação heurística. Esse grau de diferença entre as respostas é medido por meio da distância de Levenstein, definida pelo número mínimo de inserções e remoções de letras para transformar uma palavra em outra (BLACK, 1999). Por exemplo, a distância de levenstein 'ada y' e 'a day' é 2, pois pode-se realizar mínimo duas transformações: 'a day' para 'a da y' e para 'a day'.

Deste modo, no gráfico 1, representa hashtags, em que a distância de Levenstein entre a segmentação do GPT-2 e a do método heurístico é igual a 1, $d(\text{GPT-2}, \text{Heurística}) = 1$, o gráfico 2 agrupa estatísticas referentes a uma distância de repostas igual a 2, $d(\text{GPT-2}, \text{Heurística}) = 2$, e assim por diante. Por questões meramente estéticas, o gráfico 5, excepcionalmente, agrupa estatísticas referentes a distâncias de edição iguais ou maiores do que 5, $d(\text{GPT-2}, \text{Heurística}) \geq 5$,.

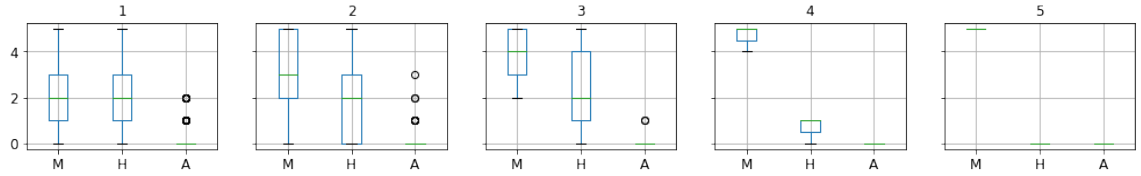

Figura 3: Gráfico com boxplots que demonstram o desempenho do modelo GPT-2 (M), segmentação heurística (H) e anotador humano espontâneo (A) para diferentes níveis de distância de edição (de 1 a 5) entre as segmentações propostas pelo GPT-2 e pela segmentação heurística.

Cada boxplot representa a distância entre cada anotador e o *gold standard*, construído pelo ser humano que realizou segmentações com consulta:

- boxplot **M**[3], distância entre o método GPT-2 e o *gold standard*.
- boxplot **H**, distância entre a heurística e o *gold standard*.
- boxplot **A**, distância entre o anotador humano espontâneo e o *gold standard*.

Observando os gráficos, constatamos que, quanto maior a distância de edição entre a segmentação heurística e a GPT-2, mais essas abordagens se diferenciariam. A GPT-2 maximiza a sua distância do *gold standard*, enquanto a segmentação heurística converge para valores mínimos. Em outras palavras, isso significa que, quanto maior for a distância de edição entre as segmentações propostas pela GPT-2 e a segmentação heurística, maior é a chance de que a segmentação heurística esteja correta e a GPT-2 esteja errada.

Na maior parte das segmentações, o modelo GPT-2 e a segmentação heurística não apresentam grande distância entre si. Isso explica porque o coeficiente kappa, referenciado na subseção

[3] Não utilizou-se a sigla G para não haver confusão entre *gold standard* e GPT-2

6.1, atribui vantagem ao modelo GPT-2, dado que é esperado que ele apresente melhores resultados na maioria dos casos. Porém, o que a Figura 3 demonstra é que, para a minoria dos casos onde a segmentação heurística discorda amplamente do GPT-2, com larga distância de edição, é mais provável que se trate de uma anomalia do modelo GPT-2, e que a segmentação heurística esteja mais próxima do *gold standard*.

7 Conclusões

Muito embora o nosso trabalho tenha conseguido atingir o objetivo de produzir uma quantidade considerável de *hashtags* segmentadas em um período computacionalmente viável, percebemos que lidar com um alto volume de dados foi uma das principais problemáticas de nossa pesquisa. Para trabalhos futuros, faz-se necessário pesquisar e considerar abordagens otimizadas que permitam lidar com um maior volume de dados em menos tempo, permitindo abranger uma janela temporal maior durante a segmentação de *hashtags*.

No campo dos resultados, demonstramos que existe uma necessidade de integração complementar entre abordagens de *deep learning* (representadas pela GPT-2) e abordagens estatísticas (representadas pela segmentação heurística) para atingir os melhores resultados possíveis na tarefa de segmentação de *hashtags*. Conclui-se que informações colhidas de um dataset apropriado de *hashtags* heuristicamente segmentadas pode auxiliar na contenção de anomalias produzidas por modelos de linguagem.

Sendo assim, é visado para trabalhos futuros a construção de datasets maiores, em mais idiomas, através de *pipelines* otimizados, e também realizar a avaliação de modelos para segmentação de *hashtags* através de experimentos que levam em conta a variação temporal que ocorre no contexto de redes sociais.

Agradecimentos

Esta pesquisa foi possível graças aos recursos computacionais do LaMCAD/UFG.

Referências

BANERJEE, Mousumi et al. Beyond kappa: A review of interrater agreement measures. **Canadian journal of statistics**, Wiley Online Library, v. 27, n. 1, p. 3–23, 1999.

BLACK, Paul E. Levenshtein distance. In: DICTIONARY of Algorithms and Data Structures [online]. 15 May 2019. [S.l.]: CRC Press LLC, 1999. (Algorithms and Theory of Computation Handbook). (accessed 14 Nov 2020) Available from: www.nist.gov/dads/HTML/Levenshtein.html.

ÇELEBI, Arda; ÖZGÜR, Arzucan. Segmenting Hashtags using Automatically Created Training Data. In: PROCEEDINGS of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), mai. 2016. p. 2981–2985. Disponível em: ¡https://www.aclweb.org/anthology/L16-1476¿. Acesso em: 27 set. 2020.

DOVAL, Yerai; GÓMEZ-RODRÍGUEZ, Carlos. Comparing neural- and N-gram-based language models for word segmentation. **Journal of the Association for Information Science and Technology**, Wiley, v. 70, n. 2, p. 187–197, 2018. ISSN 2330-1635. DOI: 10.1002/asi.24082. Disponível em: ¡http://dx.doi.org/10.1002/asi.24082¿.

MADDELA, Mounica; XU, Wei; PREOŢIUC-PIETRO, Daniel. Multi-task Pairwise Neural Ranking for Hashtag Segmentation. **arXiv:1906.00790 [cs]**, jun. 2019. arXiv: 1906.00790. Disponível em: ¡http://arxiv.org/abs/1906.00790¿. Acesso em: 29 set. 2020.

RADFORD, Alec et al. Language models are unsupervised multitask learners. **OpenAI Blog**, v. 1, n. 8, p. 9, 2019.

VASWANI, Ashish et al. **Attention Is All You Need**. [S.l.: s.n.], 2017. arXiv: 1706.03762 [cs.CL].

YANG, Jaewon; LESKOVEC, Jure. Patterns of temporal variation in online media. en. In: PROCEEDINGS of the fourth ACM international conference on Web search and data mining - WSDM '11. Hong Kong, China: ACM Press, 2011. p. 177. ISBN 978-1-4503-0493-1. DOI: 10.1145/1935826.1935863. Disponível em: ¡http://portal.acm.org/citation.cfm?doid=1935826.1935863¿. Acesso em: 27 set. 2020.

ZHAO, Hai et al. Chinese word segmentation: Another decade review (2007-2017). **arXiv preprint arXiv:1901.06079**, 2019.