Embeddings de Parágrafos de Cartas Escritas por Virginia Woolf Para Análise de Sentimento

Júlia Figueiredo S. Falcão

Instituto de Computação
Universidade Federal Fluminense
juliafalcao@id.uff.br

Abstract

Esse trabalho avalia a possibilidade de representar textos de cartas a nível de parágrafos, utilizando *embeddings* gerados pelo modelo *Word2Vec* de palavras selecionadas do parágrafo para representá-lo. Os textos passaram por estágios de pré-processamento, tokenização e lematização e a qualidade dos *embeddings* foi avaliada através da clusterização dos parágrafos, objetivando a análise de sentimento não-supervisionada. O estudo foi feito com cartas escritas pela autora modernista inglesa Virginia Woolf.

1 Introdução

Aos 9 anos, Virginia Woolf já escrevia uma revista de notícias para sua família, vivendo no bairro londrino de Kensington. Assim acredita-se ter iniciado a carreira prolífica da escritora, considerada uma das figuras mais proeminentes da literatura do século XX. Como romancista, produziu contribuições importantes para o modernismo, e além da ficção, também publicou diversos ensaios, muitos destes sobre a relação entre a mulher, a escrita e a sociedade. Ela e o marido eram donos de uma editora, responsável pela publicação de obras de diversas outras autorias famosas.

Ela também mantinha o hábito da escrita em sua vida pessoal, na forma de diários que cobrem décadas de sua vida, e frequentes cartas para seus amigos, nas quais conversava sobre os mais diversos assuntos. Era uma grande observadora da vida cotidiana.

Ao longo de tudo isso, entretanto, Virginia batalhava sintomas graves de transtorno mental, causado ou intensificado por traumas familiares e uma série de perdas trágicas. Enfrentou várias crises, algumas culminando em tentativas de suicídio. Em seus períodos mais depressivos, não conseguia ler ou escrever, o que lhe causava frustração imensa. "Estou certa de que estou enlouquecendo novamente", disse na carta que deixou para seu marido antes de se afogar no rio próximo à sua casa. Acadêmicos da literatura e da psicologia acreditam que Virginia sofria de transtorno bipolar [Boeira *et al.*, 2017], devido aos registros de seus períodos de intensa depressão, alternados com fases em que demonstrava comportamento maníaco.

Sua escrita profissional é marcada por essa batalha, tendo um aspecto sombrio, sério e até torturado às vezes. Entretanto, a pergunta que motiva este trabalho é: o quanto disso transparecia em sua escrita pessoal? Utilizando técnicas de aprendizado de representações, visamos encontrar uma maneira de representar as cartas que possa ser utilizada para efetuar uma análise de sentimento, e captar alterações de humor e períodos mais conturbados, através do modo como a escritora se comunicava com as pessoas próximas a ela. O foco está na construção dessas representações do texto, conhecidas como *embeddings*, cuja qualidade é essencial para o melhor desempenho das técnicas de aprendizado de máquina usadas no estágio seguinte.

De modo a lidar com a grande variância nos tamanhos das cartas, e melhor representar as variações de assunto e humor ao longo de uma única carta, as representações foram construídas ao nível de parágrafos. Os *embeddings* de parágrafos são constituídos de *embeddings* de palavras selecionadas. As técnicas utilizadas nesse processo serão detalhadas nas seções seguintes.

Uma vez gerados os *embeddings* de parágrafos, a análise de sentimento foi efetuada de maneira não-supervisionada, pois não há anotação a respeito do que se deseja analisar nas cartas. Foi utilizado o algoritmo de clusterização *K-Means* para agrupar os parágrafos em categorias e analisar informalmente as diferenças entre cada uma.

Na seção 2, introduzimos a origem dos dados utilizados. Nas seções 3, 4 e 5, detalhamos todo o processo de preparação dos textos, geração de *embeddings* e avaliação. Concluímos o trabalho na seção 6 com a proposta de ideias para dar continuidade a este projeto.

2 Trabalhos Relacionados

A motivação deste trabalho é similar a de trabalhos que buscam indícios de transtornos mentais, como depressão, em textos pessoais publicados na Internet. A maioria destes trabalha com textos de redes sociais, como Twitter [Shen *et al.*, 2017] e Facebook, onde as publicações costumam ser curtas e apresentar características peculiares da comunicação na Internet. Nestes domínios, os usuários tendem a se expressar de maneira informal e utilizar amplamente recursos como abreviações, sarcasmo, pontuação diferenciada e símbolos como *emojis*. Alguns desses aspectos também estão presentes nas cartas de Virginia Woolf, que escrevia de maneira mais

informal com seus amigos em correspondência do que em seus livros publicados, mas a linguagem na época era bastante diferente da linguagem informal usada hoje em dia.

Tratando-se especificamente da Virginia Woolf, o trabalho de [de Ávila Berni *et al.*, 2018] avaliou a possibilidade de utilizar ferramentas de classificação de texto para identificar indícios de suicídio. Os autores, especialistas no estudo de transtorno bipolar, compararam textos de cartas e diários escritos em períodos aleatórios de sua vida com textos dos dois meses anteriores a sua morte, e obtiveram acurácia em torno de 80% na classificação.

O trabalho de [Esposito *et al.*, 2016] comparou técnicas para a tarefa de modelagem de tópicos, uma delas sendo a clusterização de *embeddings Word2Vec*, e a outra, LDA (Latent Dirichlet allocation). Os resultados obtidos com a clusterização de *embeddings Word2Vec* foram melhores do que com LDA, ambos utilizando técnicas de pré-processamento de texto antes.

3 Dados

Muitas das cartas escritas por Virginia Woolf foram publicadas após sua morte, por Nigel Nicolson, filho de Vita Sackville-West e Harold Nicolson, que eram amigos próximos de Virginia e seu marido Leonard Woolf. A coletânea, intitulada *The Letters of Virginia Woolf*, possui seis volumes e sua versão digitalizada foi a fonte dos dados deste trabalho.

Os volumes foram obtidos em formato de livro eletrônico ePUB, e a ferramenta Calibre[1] foi utilizada para convertê-los em arquivos HTML. Destes arquivos, o texto e os dados de cada carta (como remetente e data) foram extraídos com a biblioteca BeautifulSoup[2] para Python.

No total, foram obtidas 3.756 cartas enviadas ao longo de 45 anos, cobrindo o período de 1896 até 1941, para mais de 250 destinatários diferentes. São, em média, 83 cartas por ano. As cartas têm uma média de 285 palavras (considerando os textos brutos, antes de qualquer pré-processamento), e a quantidade média de parágrafos é 5.

Além das cartas, foram utilizados alguns livros[3] da escritora para treinar o modelo *Word2Vec* junto às cartas, o que será detalhado na próxima seção. Os livros foram obtidos digitalmente da mesma maneira que as cartas.

4 Metodologia

O processo desde a obtenção dos textos brutos até a clusterização foi dividido em estágios. A cada estágio, os textos transformados foram armazenados em tabelas em arquivos CSV ou JSON, caso houvesse a necessidade de recuperá-los.

[1] www.calibre-ebook.com

[2] www.crummy.com/software/BeautifulSoup/bs4/doc

[3] *The Voyage Out, Kew Gardens, Night and Day, Monday or Tuesday Jacob's Room, Mr. Bennett and Mrs. Brown, Mrs. Dalloway, To The Lighthouse, Orlando: A Biography, The Waves, The Years, Three Guineas, Between The Acts*

Todo o procedimento descrito neste trabalho foi escrito em linguagem Python. As bibliotecas Numpy[4] e Pandas[5] foram usadas para armazenar e manipular os dados na forma de tabelas e vetores, e efetuar cálculos necessários.

4.1 Preparação ou pré-processamento

As datas das cartas encontram-se listadas nos livros em formatos variados, como "6th June [1912]" ou "2nd Feb 39". Foi mantido somente o ano, utilizando uma função personalizada com auxílio da biblioteca dateutil[6] para extraí-lo das datas, e completar as datas escritas em dois dígitos (como "39") com "18" ou "19".

Nos textos da cartas foram efetuadas diversas transformações para deixá-los no formato mais padronizado possível para o próximo estágio. No processo de extrair os textos dos arquivos HTML, cada início de parágrafo foi marcado com o símbolo "§". Foram removidos todos os outros símbolos, e os pontos final, de exclamação e de interrogação foram substituídos por um símbolo *unicode* de círculo para demarcar as sentenças, o que será necessário mais para a frente. A expressão regular usada para substituição dos pontos foi feita para desconsiderar pontos usados em acrônimos, como "V.W.". Alguns outros símbolos de pontuação, como hífens e travessões, foram substituídos por espaços, e o resto dos símbolos não-alfabéticos, como números, foram removidos do texto, assim como quebras de linha.

Em seguida, o texto foi tokenizado, ou convertido em uma lista de palavras (*tokens*). Das listas de tokens foram removidas as chamadas *stop words*, palavras que aparecem com frequência alta demais para trazer alguma informação relevante, como pronomes e artigos. A lista foi obtida da biblioteca NLTK[7], que possui ferramentas de processamento de linguagem natural, e modificada para incluir outros termos encontrados no *corpus* de cartas.

Por fim, foi feito um processo de lematização do texto, ou seja, palavras flexionadas (por exemplo, *saying*) foram substituídas por seus *lemas* (*say*), de modo que diversas variações de um mesmo termo sejam representadas no mesmo modo. O *WordNetLemmatizer* da biblioteca NLTK foi utilizado, em conjunto com uma ferramenta da mesma biblioteca para rotulação POS (*part-of-speech*). Vale ressaltar que as funções utilizadas não foram personalizadas ou aperfeiçoadas para essa base de dados específica, e não foi o objetivo conseguir um resultado ótimo de lematização, mas sim bom o suficiente para prosseguir.

O texto foi dividido em parágrafos, e cada parágrafo, dividido em sentenças. Os parágrafos foram guardados em uma tabela diferente, contendo, para cada um, o identificador da carta a qual pertence e um *offset* (igual a 0 para o primeiro parágrafo da carta e assim por diante). Cada parágrafo foi guardado na forma de uma lista de sentenças, cada sentença sendo uma lista de *tokens*. Essa tabela de parágrafos tokenizados é a utilizada no próximo estágio.

[4] www.numpy.org

[5] pandas.pydata.org

[6] dateutil.readthedocs.io

[7] nltk.org

4.2 Geração de *embeddings* de palavras

A motivação para utilizar *embeddings* a nível de parágrafos foi o desejo de preservar as variações de assunto e de humor ao longo de uma mesma carta. Virginia Woolf, conhecida por revolucionar uma técnica de escrita chamada "fluxo de consciência", na qual o autor busca transcrever o complexo processo de pensamento do personagem com associações de pensamentos e raciocínios lógicos intercalados com impressões pessoais, também faz algo similar em sua escrita pessoal. Muitas vezes os parágrafos seguem linhas de raciocínio confusas, falando de mais de um tópico ao mesmo tempo, com frases longas e reflexões complexas.

Assim, não pareceu ideal utilizar um algoritmo como *Doc2Vec* [Le and Mikolov, 2014] para representar uma carta inteira, ou mesmo um parágrafo inteiro. Em experimentos preliminares com esse algoritmo, a qualidade dos *embeddings* obtidos deixou a desejar.

No entanto, também não é possível representar um parágrafo através dos *embeddings* de todas as suas palavras, devido ao tamanho variável dos parágrafos. Ainda assim, optamos por construir primeiro os *embeddings* a nível de palavras.

Para isso, utilizamos o modelo de *Word2Vec*. Proposto por [Mikolov *et al.*, 2013], esse modelo é baseado na ideia de que se pode inferir o significado de uma palavra por sua "companhia", ou seja, o contexto, formado pelas palavras ao redor dela. Ao contrário de modelos como *Bag-of-Words*, o *Word2Vec* leva em consideração a ordem das palavras em uma frase.

O modelo é implementado pela biblioteca Gensim[8] para Python, na qual foi portado da implementação original[9] em C, e é possível treiná-lo de maneira eficiente possuindo um compilador C. Ele gera representações vetoriais densas na dimensão desejada, e para esse *corpus*, os melhores resultados foram obtidos utilizando vetores de dimensão 250. Esse valor foi escolhido através de experimentos simples e comparação dos resultados de termos mais similares a um determinado termo. Por exemplo, para a palavra *book*, a versão final do modelo retorna como mais similares as palavras *novel*, *ms* (abreviação de *manuscript*), *magazine*, *edition* e *memoir*, todas essas sendo palavras que são usadas com frequência no mesmo contexto que *book*.

O *Word2Vec* deve ser treinado com sentenças, que são de onde ele obtém o contexto de uma palavra. Por essa razão, mantivemos a separação entre as frases e guardamos os parágrafos como listas de sentenças. Para treinar o modelo com mais textos do que somente as cartas e obter um vocabulário maior, utilizamos também textos de livros da escritora. Estes passaram pelo mesmo procedimento de pré-processamento que as cartas, resultando em listas de sentenças, que são listas de palavras tokenizadas e lemmatizadas. O treinamento foi feito em 45 épocas.

4.3 Construção de *embeddings* de parágrafos

Os parágrafos resultantes após todos os passos de pré-processamento têm, em média, 22 palavras. Para representá-los, foram usados os *embeddings* de palavras obtidos pelo Word2Vec. A quantidade foi fixada em 10 palavras por parágrafo, e os parágrafos com menos ou somente 10 palavras foram desconsiderados. Os *embeddings* de palavras foram concatenados para formar o *embedding* do parágrafo, gerando um vetor de dimensão 2.500.

Para selecionar quais palavras seriam utilizadas na representação, calculamos a medida de Tf–idf (*term frequency–inverse document frequency*). Utilizamos o *TfIdfVectorizer* da biblioteca scikit-learn[10].

Essa media serviu para determinar a frequência de cada palavra em cada parágrafo, ponderada pela frequência dessa palavra em todo o *corpus*, de modo a valorizar palavras que aparecem com menos frequência, ou seja, diferenciam um parágrafo dos outros. O Tf–idf de cada palavra foi calculado e guardado na tabela para cada respectivo parágrafo.

As palavras de Tf–idf maior são as palavras mais "raras" presentes em cada parágrafo, o que significa que, como não são encontradas muitas vezes no *corpus*, os *embeddings* foram gerados com poucos exemplos de contexto. As palavras de Tf–idf menor, por outro lado, são bastante frequentes, geralmente verbos como *think* e *read*, e portanto, não agregam muita informação sobre um parágrafo específico. Assim, optamos por utilizar os valores medianos de Tf–idf para escolher as 10 palavras de cada *embedding* de parágrafo.

O procedimento de seleção das palavras foi:
Para cada parágrafo:

1. Construir uma lista de palavras (W) e outra lista (S) de seus respectivos valores Tf–idf. Construir uma lista vazia (E) de palavras selecionadas.
2. Calcular a média (m) dos valores de S.
3. Repetir 10 vezes:
   (a) Encontrar o elemento *S[i]* que tem a menor diferença relativa a m em módulo, e guardar seu índice i.
   (b) Adicionar à lista E a palavra *W[i]*.
   (c) Remover das listas W e S os elementos *W[i]* e *S[i]*.

   Em cada iteração seguinte, será encontrado mais uma palavra com Tf–idf mais próximo da média, e ao fim, obtivemos 10 palavras de Tf–idf mediano.

Por fim, os *embeddings* dessas palavras são obtidos com o modelo *Word2Vec* já treinado, e concatenados.

5 Avaliação

O método escolhido para avaliar a qualidade dos *embeddings* foi a clusterização dos mesmos, de modo a verificar a formação de grupos que reuníssem parágrafos similares em tom ou em assunto.

O algoritmo usado foi o *K-Means* [Macqueen, 1967], que é um método heurístico de clusterização. Ele funciona da seguinte maneira: São selecionados k centróides aleatórios, e então subsequentes iterações visam otimizar as posições dos centróides até que (a) não haja mais alteração em suas posições, ou (b) o número máximo estabelecido de iterações

[8] radimrehurek.com/gensim

[9] code.google.com/archive/p/word2vec

[10] scikit-learn.org

seja atingido. Esses centróides serão os centros de cada *cluster*, e cada novo vetor será alocado ao *cluster* do qual estiver mais próximo.

Utilizamos a implementação do *K-Means* da biblioteca NLTK. Tanto o número de *clusters* quanto a métrica de distância são passados como parâmetros, e optamos por dividir em 6 *clusters* e usar a métrica de similaridade de cossenos, por ser uma métrica popularmente usada para similaridade de *embeddings* gerados pelo *Word2Vec*. Além disso, para escolher as posições iniciais dos centróides, passamos um gerador de números aleatórios com um valor de semente para possibilitar a reprodução dos resultados.

O algoritmo, inicializado com os parâmetros mencionados, foi utilizado para agrupar os parágrafos resultantes dos estágios de pré-processamento descritos na seção anterior. Armazenamos na tabela o grupo atribuído a cada parágrafo.

De modo a verificar como os *embeddings* foram formados, guardamos na tabela de parágrafos as 10 palavras que foram selecionadas para representar cada um. Para visualizar o contexto de cada *cluster*, selecionamos uma amostra aleatória de 10 parágrafos de cada um.

Essas são algumas das palavras encontradas nos *embeddings* agrupados em cada um dos *clusters*:

- Cluster 0: *bad, mare, brave, unstinted, suspect, tea, dead, music, sincere, dry, hot, cypresses, suicide, hope*
- Cluster 1: *dirt, food, ruin, smoking, lapse, unexploded, remains, smash, bomb, creep, aimlessly, crank, cigar, irregularity, expression, drug, headache, heart, pest, snake, nervous, selfishness, bed*
- Cluster 2: *slowness, irregularity, tactful, loss, sardonic, apprehensive, fawn, wallow, unfairly, bewilder, shy, glasshouse, sorrow, vehement, disagree, disappointment*
- Cluster 3: *marry, happy, mind, gifted, peacefully, satisfied, ascend, hospitality, extraordinary, die, drowsy, toast, unhelped, excited, greatly, surly, womanhood, spite*
- Cluster 4: *enquiry, modern, literature, american, printing, press, poem, pleasure, thank, silly, drain, selfish, novelist, critic, story, dissertation, letter, sentence*
- Cluster 5: *finish, cope, needle, act, explain, tuberculosis, worried, rejuvenate, salvage, question, weigh, bother, autobiography, journey, gossip, weekend, humour, party*

Os *clusters* 2, 3 e 5 incluíram parágrafos de tom mais sério, como por exemplo, Virginia falando sobre sua casa em Londres ter sido destruída por uma bomba alemã durante a Segunda Guerra Mundial, e sobre doenças e preocupações. O *cluster* 4 reuniu parágrafos com termos diversos relacionados à literatura, no qual ela fala sobre seu trabalho e sobre escrever críticas, romances, dissertações e cartas.

No entanto, os *clusters* não ficaram tão bem definidos quanto o esperado.

6 Conclusão

Este trabalho abordou a possibilidade de utilizar *embeddings* de palavras selecionadas para representar parágrafos de cartas, e lidar com as variações de assunto ao longo de um mesmo parágrafo. As palavras foram selecionadas usando o valor médio de Tf–idf. A qualidade dos *embeddings* gerados foi avaliado com uma clusterização simples usando o método *K-Means*, e os resultados obtidos não demonstraram grande diferença entre um *cluster* e outro.

Para dar continuidade a este projeto, temos as seguintes ideias:

- Reavaliar a quantidade de palavras usadas para representar cada parágrafo, e qual quantidade maior pode ser escolhida de modo que não sejam perdidos muitos parágrafos que tenham menos palavras do que essa quantidade.
- Experimentar com diferentes maneiras de unir os *embeddings* de palavras, além da concatenação. Aprofundar as tarefas usadas para avaliar a qualidade dos *embeddings*, focando na análise de sentimento para conseguir um método que seja capaz de detectar as variações de humor nos textos.

Todo o código está disponível abertamente no GitHub[11].

References

[Boeira *et al.*, 2017] Manuela V. Boeira, Gabriela de Á. Berni, Ives C. Passos, Márcia Kauer-SantAnna, and Flávio Kapczinski. Virginia Woolf, neuroprogression, and bipolar disorder. *Brazilian Journal of Psychiatry*, 39:69 – 71, 03 2017.

[de Ávila Berni *et al.*, 2018] Gabriela de Ávila Berni, Francisco Diego Rabelo-da Ponte, Diego Librenza-Garcia, Manuela V. Boeira, Márcia Kauer-Sant'Anna, Ives Cavalcante Passos, and Flávio Kapczinski. Potential use of text classification tools as signatures of suicidal behavior: A proof-of-concept study using virginia woolf's personal writings. *PLOS ONE*, 13(10):1–11, 10 2018.

[Esposito *et al.*, 2016] Fabrizio Esposito, Anna Corazza, and Francesco Cutugno. Topic modelling with word embeddings. 12 2016.

[Le and Mikolov, 2014] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

[Macqueen, 1967] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[Shen *et al.*, 2017] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression detection via harvesting social

[11] github.com/juliafalcao/yours-ever

media: A multimodal dictionary learning solution. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3838–3844, 2017.