Quem é Estamira? Uma análise da coerência dos discursos através de combinação de medidas para classificação de texto

Bruno Aristimunha[1]

[1]Universidade Federal do ABC – Santo André – SP – Brasil.

Abstract. *This article investigates disorders in natural speech through natural language processing techniques. We analyze the possible incoherent/schizophrenic speech using the Latent Semantic Analysis method and connectivity measures of the co-occurrence graph. Using these measures as attributes, we selected* 4 *classifiers to distinguish illogical speech patterns. The chosen corpus comes from the lines of characters from the film Estamira (2004), which tells the life of a garbage collector from Rio de Janeiro. This choice is justified by the fact that the narrative is anchored in complex talks of its main character, Estamira, who for long periods conducts obscure dialogues. There is also a discursive clash between the character's lines and the eloquence of her family.*

Resumo. *Este artigo investiga os distúrbios no discurso natural através de técnicas de processamento de linguagem natural. Analisamos o possível discurso incoerente/esquizofrênico através do método de Análise Semântica Latente e de medidas de conectividade do grafo de co-ocorrência. Combinando essas medidas como atributos, empregamos* 4 *classificadores para distinguir padrões de fala ilógicos. O corpus escolhido é oriundo das falas de personagens do filme Estamira (2004), que conta a vida de uma catadora de lixo do Rio de Janeiro. A escolha da obra justifica-se por uma narrativa ancorada em discursos complexos de sua personagem principal, Estamira, que durante longos períodos conduz diálogos abstrusos. Há, também, choque discursivo entre as falas da personagem e a lógica na eloquência de seus familiares*[*].

1. Introdução

Esquizofrenia é um transtorno psicótico grave de consequência agressiva para os indivíduos. A doença possui alta correlação com situações de vulnerabilidade social e outras patologias neurológicas, como falta de moradia, depressão, ansiedade, American-Psychiatric (2013). O espectro da esquizofrenia é diagnosticado pela presença de sintomas como delírios, alucinações, pensamento desorganizado (discurso) e comportamento motor grosseiro.

O transtorno do neurodesenvolvimento possui sinal claro na linguagem produzida pelo paciente soando extremamente inteligível. Um exemplo de frase inteligível vem do linguista Chomsky (1956) com a frase "ideias verdes incolores dormem furiosamente", que apesar de sintaticamente correta, não apresenta semântica lógica alguma. O diagnóstico prematuro dos sintomas cardinais é de supra necessidade para o tratamento e amenização das fases agudas da psicose. A característica do pensamento desorganizado e incoerência semântica se mostra preditora pelo menos 3 anos antes do primeiro surto psicótico, Bedi et al. (2015).

A grande problemática da avaliação do discurso é que as pessoas que sofrem estresse emocional são incapazes de reconhecer os sintomas, Peralta e Cuesta (1998). Para contornar essa problemática e distinguir pacientes com esquizofrenia, nessa pesquisa pretende-se avaliar um espaço de fala de uma pessoa esquizofrênica no filme Estamira. Esse filme é considerando um dos mais fieis com a realidade esquizofrênica, Ventura (2008). A obra retrata de forma semi-estruturada diversos aspectos da vida de Estamira, e possui breves espaços de fala de pessoas com pensamento organizado. Pesquisas anteriores já demonstraram sucesso no reconhecimento de padrões de fala de pessoas esquizofrênicas. Dentre essas pesquisas, podemos citar algumas características

[*]Código em: estamira-coerencia-discursos, em qualquer interesse escreva: b.aristimunha@gmail.com

importantes no discurso esquizofrênico: alto uso de pronomes de auto-referencialmente e avolia tendendo aos estados negativos, Buck et al. (2015). Alguns estudos já colaboram com aplicações de grafos para avaliar conexões de forma automáticas através da fala, Mota (2012). Assim, o objetivo desse trabalho é analisar e classificar de forma automática a coerência do discurso na esquizofrenia. Para tanto, devido uma escassez de corpus de pacientes esquizofrênicos selecionamos o filme Estamira.

2. Fundamentos

2.1. Matriz de co-ocorrência

O texto a ser analisado é representado por uma matriz de co-ocorrência $\mathcal{M}$, em que por linha temos uma palavra única no texto e por coluna temos a ocorrência das n palavras adjacentes. Entendemos como adjacente toda palavra vizinha a essa no texto. Através do n determinamos a profundidade da busca da relação. Quanto maior o nosso n, mais semântica será nossa análise, quando menor, mais sintática Jurafsky e Martin (2009).

2.2. Análise Semântica Latente

Análise Semântica Latente - *ASL* consiste em uma abordagem matemática e computacional para descoberta das relações de semelhanças entre textos, excertos ou sentenças dentro de um texto (distribuição semântica). A premissa do método baseia-se em assumir que palavras usadas em contextos similares tendem a ser semanticamente mais semelhantes entre si do que outras.

Com base nessa premissa, dado uma matriz esparsa de co-ocorrência ($\mathcal{M}$), aplicamos o método de decomposição em valores singulares em que obtemos um espaço semântico. Essa decomposição em valores singulares pode ser entendida como: dado uma matriz $\mathcal{M} \in \mathbb{R}^+$ com dimensões $i \times j$, fatoremos essa M na forma: $U \cdot \Sigma \cdot V^*$, que U é uma matriz unitária ($i \times j$), Σ é matriz com valores singulares sobre a diagonal ($i \times j$), e V^* sendo a conjugada transposta de V, uma matriz unitária do produto entre U e V^*, que por sua vez produz uma matriz identidade.

Com essa redução induzimos similaridades semânticas da linguagem com base no padrão de uso das palavras em corpus. Através de uma heurística pode-se também descartar as k colunas finais da matriz U, e a matriz Σ e V. A matriz U agora com dimensões $m \times k$ possui uma densidade maior se comparada com a matriz M. Esse espaço semântico, dado pela soma dos vetores das palavras individuais, permite medir a quantidade de similaridade semântica.

2.3. Modelagem por grafo

Uma outra abordagem possível para analisar diferentes aspectos da linguagem pode ser feita através de modelagem de grafos. Um Grafo $\mathcal{G}$ é definido como um par ordenado $\langle \mathcal{V}, \mathcal{E} \rangle$, em que $\mathcal{V}$ é conjunto finito e não-vazio de vértices (também conhecido como nós) e $\mathcal{E}$ é o conjunto de arestas (também conhecido como links) dos vértices, $\{\mathcal{E} \subseteq (u,l) | \{u,l\} \in \mathcal{V}\}$, se (u,l) é uma aresta, logo, os vértices u e l são adjacentes. Desse grafo $\mathcal{G}$, inferimos medidas globais sobre os padrões presentes, para descrever propriedades estatísticas nessa estrutura de dados. Para esse trabalho selecionamos duas medidas para analisar o seu comportamento, sendo elas: coeficiente de agrupamento e comprimento mínimo do caminho

O coeficiente de agrupamento quantifica a influência que um dado vértice v possui para permitir a criação de uma conexão entre os vértices vizinhos. Essa medida pode ser definida como $C_v = \frac{e_{uk}}{d_v(d_v-1)}$, em $|e_{uk}|$ é o número de dois vizinhos de um vértice v que são conectados entre si, d_v é o grau desse vértice. Computamos a medida em todos vértices e extraímos a média, como: $C(\mathcal{G}) = \frac{1}{N} \sum_{v=1}^{N} C_v$, em que $N = |V|$.

O comprimento mínimo do caminho pode ser definido como: dado um grafo não direcionado $\mathcal{G}$ com um conjunto de vértices $\mathcal{V}$, com uma função peso tal que $w : \mathcal{E} \rightarrow \mathbb{R}$, mapeando

as arestas para valores reais. Definimos como peso do caminho do curto $\delta(u, l)$ desde u até l por: $\delta(u, l) = min\{w(p) : u \rightarrow^p l\}$, se há um caminho entre, do contrário 0. O caminho mais curto será desde vértice i até vértice j e então definido como qualquer caminho p com peso $w(p) = \delta(u, l)$. Para o caminho mais curto para todos os pares de vértice, soma-se tudo e se divide pelo número de pares de vértices $(N \cdot (N-1))$, ou seja, $L(G) = \frac{1}{N \cdot (N-1)} \sum_{u \neq l} d(v_u, v_l)$. Como o grafo representa um recorte da fala do filme Estamira, usamos essas medidas para quantificar as diferenças estruturais do discurso de Estamira e outros personagens ("não Estamira").

3. Materiais e Métodos

Reconhecimento do discurso desorganizado geralmente envolve uma análise de como o indivíduo fala em entrevistas semiestruturadas. Como um discurso não coerente pode ser reconhecido facilmente, após um processo de rotulamento é possível a aplicação de algoritmos capazes de reconhecer esse sintoma cardial da esquizofrenia sem entrevistas semiestruturadas. A Figura 1 ilustra o fluxograma para nossa arquitetura, sendo dividia em cinco etapas: pre-processamento (rotulamento), matriz co-ocorrência, grafo, análise de coerência e classificação.

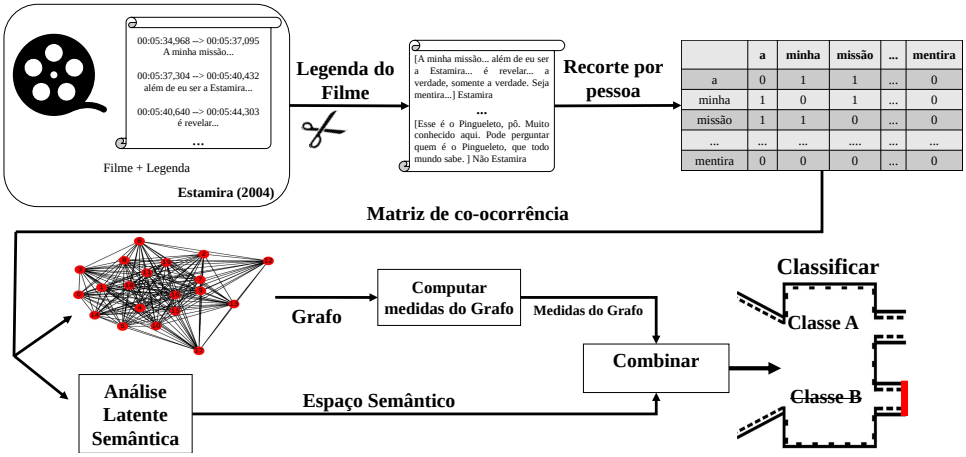

Figura 1. Cada bloco representa um procedimento e cada seta o fluxo de informação.

No processo de rotulamento realizamos o recorte das falas de Estamira. Para isso, assistimos o filme, e a cada transição de cena, mudança de assunto ou de personagem, recortamos a fala. No conjunto de falas, rotulamos essas como somente fala da personagem Estamira, e como não pertencente a Estamira ("Não Estamira"). Por fala, geramos uma matriz de co-ocorrência com $n = 79$ para análise de Grafo, e também, em paralelo, aplicamos a análise Semântica Latente, sendo o número 79 corresponde à quantidade media de palavras na frase independente da Classe. Para tarefa do conjunto de dados usado na classificação combinamos as métricas do grafo co-ocorrência, algumas informações sobre as frases, como: a quantidade de caracteres e palavras, e média e desvio padrão do primeiro componente obtido pela análise semântica latente.

Dos algoritmos mais usados na mineração de dados, segundo Wu et al. (2008), selecionamos os 4 métodos mais comum na tarefa da classificação: Máquina de Vetor de Suporte (SVM), K-vizinhos (KNN), Classificador Ingênuo Gaussiano e Árvore de Decisão nas características das medidas observadas no grafo e outras informações para predizer de quem é a fala.

4. Resultados

Para avaliar nossa metodologia, calculamos a acurácia ($\mathcal{A}$) dos quatro métodos testados, em validação cruzada com 10 folhas, com os parâmetros permaneceram com os valores padrões da biblioteca scikit-learn. Na Tabela 1 podemos observar os resultados de cada método, verificamos que todos

os classificadores apresentaram alta variância ($13\% - 34\%$), e resultados medianos, sendo o melhor SVM e o pior o K-Vizinhos.

Classificador	$\mathcal{A}$ média	$\pm$
Máquina de Suporte de Vetores	0.62	0.14
Árvores de Decisão	0.61	0.24
Naive Bayes Gaussiano	0.58	0.42
K-Vizinhos	0.44	0.36

Tabela 1. Média e Incerteza da Acurácia obtida.

Traçando paralelo com os resultados obtidos por Bedi et al. (2015), na mesma tarefa, mas com validação *leave-one-out*, nosso resultado se mostra aquém do esperado (100% de acurácia deles). Isso ocorre, em partes, por conta do ruído no nosso conjunto de dados, e também uma diferença no processo de validação. Já em Corcoran et al. (2018), os valores da acurácia variam de 72% à 83%, a depender do protocolo. Esses trabalhos dão indícios que os resultados obtidos aqui estabelecem um bom *baseline*, mas que ainda há espaço para melhorias na metodologia.

5. Conclusão

O discurso incoerente é um sintoma cardial e complexo para patologias neurológicas, que deve ser analisado e estudado com cuidado médico. Através deste trabalho conseguimos construir uma metodologia, tendo como base um contexto ruidoso de um filme, no qual distinguimos a fala incoerente de Estamira. Como próximo passo, pode-se avaliar a importância do que os classificadores estão aprendendo, visando um melhor entendimento dos padrões do discurso. Pode-se também empregar abordagens distintas para: o recorte da fala, como o reconhecimento por timbre/imagem, e extração de atributos, com modelos de linguagem, *e. g.*, BERT e FastText. Sugere-se também avaliar o contexto da detecção de anomalia, com classificação de uma classe. Propõe-se realizar uma adaptação, em uma próxima etapa, para a inclusão de atributos mais relacionados às características da doença, como contagem de pronomes e análise de sentimentos.

Agradecimentos

Aos professores, pelas sugestões durante a escolha do tema, Walter Hugo Lopez Pinaya; a escolha do *corpus*, Allan Moreira Xavier; a execução, Jesús P. Mena-Chalco. A Matheus Oliveira, pela revisão final do texto.

Referências

American-Psychiatric (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Association. doi:10.1176/APPI.BOOKS.9780890425596.

Bedi, G., Carrillo e *et. al.* (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Nature Partner Journals: Schizophrenia*, 1:15030–15037. doi:10.1038/NPJSCHZ.2015.30.

Buck, B., Minor, K. S. e Lysaker, P. H. (2015). Differential lexical correlates of social cognition and metacognition in schizophrenia; a study of spontaneously-generated life narratives. *Comprehensive psychiatry*, 58:138–145. doi:10.1016/J.COMPPSYCH.2014.12.015.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124. doi:10.1109/TIT.1956.1056813.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E. e Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75. doi:10.1002/WPS.20491.

Jurafsky, D. e Martin, J. H. (2009). *Speech and Language Processing*. Prentice Hall PTR, USA. doi:10.1162/089120100750105975.

Mota, N. B. t. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4). doi:10.1371/JOURNAL.PONE.0034928.

Peralta, V. e Cuesta, M. J. (1998). Lack of insight in mood disorders. *Affective Disorders*, 49(1):55–58. doi:10.1016/S0165-0327(97)00198-5.

Ventura, L. D. S. L. (2008). Estamira em três miradas. Mestrado em psicologia clínica e cultura, Universidade de Brasília - UNB. URL: repositorio.unb.br/handle/10482/3955.

Wu, X., Kumar, V., Ross, Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D. e Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37. doi:10.1007/S10115-007-0114-2.