

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MARCOS AUGUSTO GRZEÇA

**Drunk2Symbol e Drunk2Vec: Métodos
para a identificação de textos bêbados
explorando enriquecimento contextual**

Dissertação apresentada como requisito parcial
para a obtenção do grau de Mestre em Ciência da
Computação

Orientador: Profa. Dra. Renata Galante

Porto Alegre
2020

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Grzeça, Marcos Augusto

Drunk2Symbol e Drunk2Vec: Métodos para a identificação de textos bêbados explorando enriquecimento contextual / Marcos Augusto Grzeça. – Porto Alegre: PPGC da UFRGS, 2020.

88 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2020. Orientador: Renata Galante.

1. Classificação de textos bêbados. 2. Enriquecimento Contextual. 3. Web Semântica. 4. Word Embeddings. 5. Processamento de Linguagem Natural. I. Galante, Renata. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitor: Prof. Rui Vicente Oppermann

Vice-Reitora: Prof^a. Jane Fraga Tutikian

Pró-Reitor de Pós-Graduação: Prof. Celso Giannetti Loureiro Chaves

Diretora do Instituto de Informática: Prof^a. Carla Maria Dal Sasso Freitas

Coordenador do PPGC: Prof^a. Luciana Salete Buriol

Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

AGRADECIMENTOS

Agradeço aos meus pais Delma e Antônio, a minha irmã Francielle e a minha namorada Fernanda pelo apoio, paciência, incentivo e pelos cuidados durante esta jornada acadêmica. Todos foram de extrema importância durante os momentos de incerteza.

Agradeço a minha orientadora, Prof^a. Dr^a. Renata Galante, pela oportunidade, conselhos, zelo e ensinamentos durante os três anos. Agradeço a minha coorientadora, Prof^a. Dr^a. Karin Becker, pela grande ajuda e zelo, atuando ativamente para melhorar o trabalho e transmitir conhecimento. As minhas excelentes orientadoras prestaram todo o suporte, conhecimento, mentoria e, sobretudo, se preocuparam com o meu bem-estar. Agradeço de coração por todos os ensinamentos e pela dedicação durante esta jornada. Também agradeço aos demais professores do INF-UFRGS pelas contribuições e aos amigos que fiz durante os estudos.

Também agradeço aos meus colegas de trabalho da Tiny, que entenderam minhas ausências e deram todo o suporte necessário para auxiliar nesta jornada.

Finalmente agradeço a Deus pela oportunidade propiciada e saúde.

RESUMO

O consumo excessivo de álcool é responsável por três milhões de mortes anualmente e continua crescendo em todo o mundo, tornando-se um importante problema de saúde pública. As redes sociais provêm informações para monitorar e entender os problemas de saúde pública, inclusive o abuso de álcool. As informações extraídas das redes sociais podem auxiliar os gestores públicos a reduzir o uso nocivo do álcool, porém é necessário investir em métodos para extrair e identificar automaticamente o consumo de álcool a partir das redes sociais. Este trabalho aborda a classificação automática de textos bêbados a partir do Twitter, que consiste na classificação de tweets em {bêbado, sóbrio} de acordo com o seu conteúdo. Métodos tradicionais de processamento de linguagem natural não apresentam bom desempenho na identificação de tweets bêbados (ou seja, postados sob a influência de álcool), pois os tweets são curtos, esparsos e escritos com vocabulário específico da Internet. Para superar esses desafios e classificar os tweets, são propostos dois métodos que exploram estratégias distintas de enriquecimento contextual: *Drunk2Symbol* e *Drunk2Vec*. *Drunk2Symbol* expande o vocabulário e fornece contexto aos tweets explorando o enriquecimento contextual externo (Web Semântica). *Drunk2Symbol* também extrai *features* que caracterizam o abuso de álcool. Por outro lado, *Drunk2Vec* utiliza a semântica distribucional para identificar palavras similares e para lidar com as idiossincrasias da linguagem empregada em tweets bêbados. Para equilibrar as melhorias dos dois métodos, foi utilizado um conjunto de classificadores, denominado *Drunk2Ensemble*. Este trabalho disponibiliza duas bases de dados públicas relacionadas ao consumo de álcool e uma análise exploratória que ilustra a riqueza e a aplicabilidade das informações extraídas a partir das redes sociais. Para avaliar o desempenho dos métodos, foi definido um protocolo experimental abrangente, envolvendo três classificadores e cinco bases de dados que abordam diferentes comportamentos relacionados ao consumo de álcool no Twitter. Os resultados demonstram alto desempenho, com a medida F_1 superior a 88,8 pontos percentuais em todas as bases de dados, superando o *baseline* com melhorias estatisticamente significativas. Os métodos propostos podem identificar tweets bêbados e fornecer informações importantes que ajudam a monitorar e entender os fatores relacionados ao consumo excessivo de álcool.

Palavras-chave: Classificação de textos bêbados. Enriquecimento Contextual. Web Semântica. Word Embeddings. Processamento de Linguagem Natural.

Drunk2Symbol and Drunk2Vec: Methods to Identify Drunk Texting Exploring Contextual Enrichment

ABSTRACT

Excessive alcohol consumption causes about 3 million deaths annually and continues to grow worldwide, becoming a major public health problem. Social networks provide information to monitor and understand public health issues, including alcohol abuse. The information extracted from social network can help public managers to reduce harmful alcohol use, but it is necessary to invest in methods to automatically extract and identify alcohol consumption from social networks. This work deals with the automatic classification of drunk texting from Twitter, which consists of the classification of tweets in {drunk, sober} according to their content. Traditional methods of natural language processing do not perform well in identifying drunk tweets (i.e., posted under the influence of alcohol) because tweets are short, sparse, and written with Internet-specific vocabulary. To overcome these challenges and classify tweets, two methods that explore distinct contextual enrichment strategies are proposed: *Drunk2Symbol* and *Drunk2Vec*. *Drunk2Symbol* expands the vocabulary and provides context for tweets by exploring external contextual enrichment (Semantic Web). *Drunk2Symbol* also extracts features that characterize drunk behavior. On the other hand, *Drunk2Vec* uses distributional semantics to identify similar words and to deal with the idiosyncrasies of the language used in drunk tweets. An ensemble, namely *Drunk2Ensemble*, was used to combine the improvements of both methods. This work provides two public datasets related to alcohol consumption and an exploratory analysis that illustrate the richness and application of information extracted from social networks. To evaluate the performance of methods, a broad experimental protocol was defined, involving three classifiers and five datasets addressing different drunk texting behaviors related to alcohol consumption on Twitter. The results show high performance, with the F_1 measure higher than 88.8 percentage points in all datasets, outperforming the baseline with statistically significant improvements. The proposed methods can identify drunk tweets and provide relevant information that helps to monitor and understand factors related to excessive alcohol consumption.

Keywords: Drunk tweets classification, Contextual Enrichment, Semantic Web, Word Embeddings, Natural Language Processing.

LISTA DE ABREVIATURAS E SIGLAS

AMT	Amazon Mechanical Turk
API	Application Programming Interface
BoW	Bag of Words
CNN	Redes Neurais Convolucionais
DNN	Redes Neurais Densas
GBDT	Gradient Boosted Decision Trees
InfoGain	Information Gain
LSTM	Long Short-Term Memory
PLN	Processamento de Linguagem Natural
POS	Part of Speech
PP	Pontos Percentuais
SVM	Support Vector Machines
URL	Uniform Resource Locator

LISTA DE FIGURAS

Figura 2.1 SVM.....	16
Figura 2.2 <i>Gradient Boosted Decision Trees</i>	17
Figura 2.3 Exemplo de arquitetura da CNN para PLN	19
Figura 2.4 Codificação <i>One-hot</i> vs <i>Word embeddings</i>	20
Figura 2.5 Relacionamento entre <i>word embeddings</i>	21
Figura 2.6 CBOW ou <i>Skip-Gram</i>	22
Figura 2.7 Janela de contexto.....	22
Figura 2.8 Conjunto de classificadores com empilhamento	25
Figura 2.9 Matriz de confusão	26
Figura 4.1 Fluxograma para criação do <i>DS1-drunk</i>	38
Figura 4.2 Fluxograma para coleta do <i>DS2-keywords</i>	40
Figura 4.3 Fluxograma para criação do <i>DS3-drinking-ext</i>	42
Figura 4.4 Ferramenta para anotar <i>DS3-drinking-ext</i>	43
Figura 4.5 Nuvem de palavras para tweets sóbrios e bêbados	46
Figura 4.6 Nuvem de palavras comparando as <i>features</i> conceituais.....	47
Figura 4.7 Pirâmide demográfica	48
Figura 4.8 Tweets agrupados pela categoria da localização.....	49
Figura 4.9 Tweets organizados pelo grupo da categoria de localização	50
Figura 4.10 Sentimentos expressos nos tweets	50
Figura 5.1 <i>Features</i> extraídas usando <i>Drunk2Symbol</i> e <i>Drunk2Vec</i>	52
Figura 5.2 Visão geral do <i>Drunk2Symbol</i>	53
Figura 5.3 Exemplo de execução do método <i>Drunk2Symbol</i>	53
Figura 5.4 Visão geral do <i>Drunk2Vec</i>	58
Figura 5.5 Exemplo de execução do método <i>Drunk2Vec</i>	58
Figura 5.6 Arquitetura da CNN para aprender <i>drunk word embeddings</i>	60
Figura 6.1 Arquitetura da DNN	64
Figura 6.2 Fluxo de execução do <i>Drunk2Ensemble</i>	72
Figura 6.3 Sumarização dos resultados das variações para gerar <i>word embeddings</i>	77

LISTA DE TABELAS

Tabela 3.1 Comparativo entre os trabalhos relacionados que realizam classificação automática de textos bêbados.....	34
Tabela 4.1 Resumo das bases de dados	37
Tabela 4.2 Características do <i>DS1-drunk</i>	39
Tabela 6.1 Tamanho das bases de dados	63
Tabela 6.2 <i>Features</i> utilizadas no Experimento #1: <i>Drunk2Symbol</i>	65
Tabela 6.3 Resultados do Experimento #1	66
Tabela 6.4 Trinta <i>features</i> mais importantes para <i>Drunk2Symbol-SVM</i>	67
Tabela 6.5 Resultados do Experimento #2	70
Tabela 6.6 Melhorias obtidas usando enriquecimento contextual externo.....	71
Tabela 6.7 Diferença entre o desempenho médio do <i>Drunk2Ensemble</i> e dos classificadores individuais	73
Tabela 6.8 Comparativo entre <i>Drunk2Symbol-SVM</i> e <i>Drunk2Ensemble</i>	73
Tabela 6.9 Comparativo entre <i>Drunk2Vec-DNN</i> e <i>Drunk2Ensemble</i>	73
Tabela 6.10 Comparativo entre <i>Drunk2Vec-SVM</i> e <i>Drunk2Ensemble</i>	74
Tabela 6.11 <i>Embeddings</i> aprendidas usando CNN, <i>Word2Vec</i> e GloVe	78
Tabela 6.12 Resultados das diferentes variações de <i>word embeddings</i>	78
Tabela A.1 Índices de concordância do coeficiente <i>kappa</i>	87
Tabela B.1 DNN	88
Tabela B.2 SVM	88
Tabela B.3 XGBoost	88

SUMÁRIO

1 INTRODUÇÃO	11
2 FUNDAMENTAÇÃO TEÓRICA	15
2.1 Algoritmos de aprendizado de máquina	15
2.1.1 <i>Support Vector Machines</i>	16
2.1.2 <i>eXtreme Gradient Boosting</i>	17
2.2 Aprendizado profundo	17
2.2.1 Redes Neurais Convolucionais	18
2.2.2 Representação textual para o aprendizado profundo	19
2.3 Redução de dimensionalidade	23
2.3.1 Algoritmos de seleção de <i>features</i>	23
2.3.2 Algoritmos de extração de <i>features</i>	24
2.4 Conjunto de classificadores	24
2.5 Métricas de avaliação	25
3 TRABALHOS RELACIONADOS	27
3.1 Motivação para identificar e classificar textos bêbados	27
3.1.1 Identificação automática de textos bêbados	28
3.2 Enriquecimento contextual	29
3.2.1 Enriquecimento contextual externo	30
3.2.2 <i>Word embeddings</i>	30
3.3 Análise dos trabalhos relacionados	31
3.4 Considerações finais	33
4 BASES DE DADOS - TWEETS RELACIONADOS AO CONSUMO DE ÁL- COOL	36
4.1 Visão geral	36
4.2 Base de dados <i>DS1-drunk</i>	37
4.2.1 Coleta de dados	37
4.2.2 Anotação de dados	38
4.2.3 Dicionário de dados	39
4.3 Base de dados <i>DS2-keywords</i>	39
4.3.1 Coleta de dados	40
4.3.2 Anotação de dados e pré-processamento	41
4.3.3 Dicionário de dados	41
4.4 Base de dados <i>DS3-drinking-ext</i>	41
4.4.1 Coleta de dados	42
4.4.2 Anotação dos dados	43
4.4.3 Dicionário de dados	44
4.5 Análise exploratória dos dados	44
4.5.1 Termos mais utilizados por usuários que estão sob efeito do álcool	45
4.5.2 Pirâmide demográfica	46
4.5.3 Estabelecimentos vinculados ao consumo de álcool	47
4.5.4 Análise de sentimentos	48
4.6 Considerações finais	49
5 DRUNK2SYMBOL E DRUNK2VEC: MÉTODOS DE ENRIQUECIMENTO CONTEXTUAL PARA A CLASSIFICAÇÃO DE TEXTOS BÊBADOS ...	51
5.1 Visão geral	51
5.2 <i>Drunk2Symbol</i>	52
5.2.1 Pré-processamento e extração de features de embriaguez	52
5.2.2 Tratamento de erros	54

5.2.3	Enriquecimento contextual externo	54
5.2.4	Integração de <i>features</i>	56
5.3	<i>Drunk2Vec</i>	57
5.3.1	Pré-processamento e tratamento de erros	58
5.3.2	Representação das <i>drunk word embeddings</i>	58
5.3.3	Projeção das <i>features</i> semânticas	60
5.3.4	Integração de <i>features</i>	61
6	EXPERIMENTOS E RESULTADOS	62
6.1	Objetivos	62
6.2	Bases de dados	62
6.3	Ferramentas e métricas de avaliação	63
6.4	Experimento #1: <i>Drunk2Symbol</i>	64
6.4.1	Metodologia	64
6.4.2	Resultados	65
6.4.2.1	Desempenho na classificação de tweets bêbados	65
6.4.2.2	Contribuição das <i>features</i> propostas	67
6.5	Experimento #2: <i>Drunk2Vec</i>	68
6.5.1	Metodologia	68
6.5.2	Resultados	69
6.5.2.1	Desempenho na classificação de tweets bêbados	69
6.5.2.2	Contribuição da integração entre <i>drunk word embeddings</i> e <i>features semânticas</i>	70
6.6	Experimento #3: <i>Drunk2Ensemble</i>	71
6.6.1	Metodologia	71
6.6.2	Resultados	72
6.6.3	Casos de falhas	73
6.7	Experimento #4: CNN para gerar <i>word embeddings</i>	75
6.7.1	Metodologia	75
6.7.2	Resultados	76
6.8	Considerações finais	78
7	CONCLUSÃO	80
	REFERÊNCIAS	82
	APÊNDICE A — ÍNDICES DE CONCORDÂNCIA DO COEFICIENTE KAPPA	87
	APÊNDICE B — VARIAÇÕES DAS <i>WORD EMBEDDINGS</i> POR BASE DE DADOS	88

1 INTRODUÇÃO

O consumo excessivo de álcool está entre os maiores problemas que afetam nossa sociedade atualmente (MAITY et al., 2018). O abuso de álcool causa cerca de 3 milhões de mortes anualmente, o que corresponde a 5% de todas as mortes globais. Entre os jovens, a taxa de mortalidade relacionada ao álcool é de 13,5%. A intoxicação alcoólica também foi responsável por 15% das mortes resultantes de acidentes de trânsito (ORGANIZATION, 2019), onde os jovens são mais suscetíveis a se envolverem em acidentes fatais. O abuso de álcool também prejudica a coordenação motora, induz a comportamentos agressivos ou antissociais (PARROTT; ECKHARDT, 2018), suicídios (HEIMISDOTTIR et al., 2010) e doenças hepáticas crônicas (MAITY et al., 2018).

Compreender os fatores relacionados ao abuso de álcool é importante para permitir o desenvolvimento de políticas efetivas de saúde pública. Por exemplo, informações relacionadas às áreas mais expostas de uma cidade, bem como características demográficas dos consumidores e suas motivações, podem ajudar os gestores públicos e prestadores de serviços de saúde a tomar decisões sobre como evitar danos relacionados ao álcool. A Organização Mundial da Saúde destaca a necessidade de ferramentas para monitorar as tendências do consumo de álcool e danos relacionados. Essas ferramentas têm como objetivo fornecer, em tempo hábil, informações relevantes e confiáveis aos gestores públicos, para garantir uma avaliação eficaz das opções e intervenções públicas que podem reduzir o uso nocivo do álcool (ORGANIZATION, 2019). No entanto, coletar informações relacionadas ao consumo de álcool por meio de estudos especializados pode ser caro e demorado.

As redes sociais podem desempenhar um papel importante nesse sentido, uma vez que o consumo de álcool e o uso de redes sociais são comportamentos relacionados aos jovens. Adicionalmente, as redes sociais fornecem informações de alto valor com baixo custo e latência. A rede social Twitter é utilizada para identificar problemas de saúde pública, inclusive o consumo excessivo de álcool (CURTIS et al., 2018). Trabalhos relacionados mostram que os dados oficiais sobre o consumo de álcool e os dados obtidos por meio de tweets relacionados ao consumo de álcool estão fortemente correlacionados, confirmando a importância dos dados extraídos das redes sociais (CULOTTA, 2013; WEST et al., 2012; CURTIS et al., 2018). Também é importante destacar que a exposição repetida a postagens envolvendo drogas nas redes sociais pode incentivar outras pessoas, pois esse comportamento pode ser percebido como normal (WEST et al., 2012; JERNIGAN

et al., 2017). Portanto, o Twitter é uma fonte rica de dados em tempo real a partir do qual padrões sobre o consumo de álcool podem ser extraídos.

Conceitualmente, enviar mensagens de texto sob a influência de álcool é conhecido popularmente como *drunk texting*¹. Neste trabalho, o termo *drunk texting* é referenciado como texto bêbado. Nas redes sociais, tweet bêbado denota tweets escritos por pessoas sob influência de álcool (JOSHI et al., 2015).

Este trabalho aborda o problema da classificação automática de tweets bêbados, ou seja, dado um tweet, é necessário classificá-lo em {bêbado, sóbrio}, dependendo se o mesmo está relacionado ao consumo de álcool. A detecção de mensagens de textos bêbados é inserida na área da paralinguística e enfrenta desafios comuns à categorização de tópicos em tweets. Os tweets são curtos e os usuários empregam abreviações, *emojis*, gírias da Internet, *hashtags*, memes e URLs para transmitir suas mensagens. Além disso, o consumo de álcool afeta a capacidade de escrever mensagens corretamente, bem como as emoções expressas (BORRILL; ROSEN; SUMMERFIELD, 1987). Por exemplo, o tweet *'how do you know you had too much beer? when it runs down your chin #toohappy'* deve ser classificado como bêbado. Algumas pistas podem ser identificadas no tweet: erro de digitação (*beeer*), as expressões *too much* e *runs down to your chin* e a *hashtag* (*#toohappy*). Portanto, a identificação de tweets bêbados envolve textos esparsos, com ruídos e termos sem contexto.

Os trabalhos existentes na área limitam-se ao desenvolvimento de técnicas de aprendizado supervisionado que classificam *features*² textuais extraídas com técnicas tradicionais de processamento de linguagem natural (N-grama e *Bag-of-Words* (BoW)), sentimentos relacionados e *features* que caracterizam a morfologia da sentença para identificar o consumo de álcool (JAUCH; JAEHNE; SUENDERMANN, 2013; APHINYA-NAPHONGS et al., 2014; JOSHI et al., 2015; HOSSAIN et al., 2016; MAITY et al., 2018). Nenhum desses trabalhos tentou melhorar a classificação do texto explorando estratégias de enriquecimento contextual que possam lidar com a esparcialidade³ e o vocabulário específico utilizado em textos bêbados, nem técnicas de aprendizado profundo.

O trabalho desenvolvido nesta dissertação visa melhorar o desempenho da classificação de mensagens de textos bêbados explorando duas formas de enriquecimento contextual: *enriquecimento contextual externo* e *semântica distribucional*. O enriquecimento contextual externo generaliza *features* textuais para conceitos obtidos usando fontes exter-

¹ <https://www.urbandictionary.com/define.php?term=drunk+texting>

² Optou-se pelo termo *features*, em inglês, pela sua ampla utilização na literatura atual

³ Esparcialidade corresponde a tradução do termo *sparseness*

nas de conhecimento (por exemplo, Web Semântica), agrupando assim termos de acordo com seu significado semântico. Esse tipo de enriquecimento foi explorado com sucesso em problemas relacionados à classificação de eventos nas redes sociais (por exemplo, Schulz, Guckelsberger e Janssen (2017), Romero e Becker (2019)), mas apresenta limitações para lidar com as especificidades do vocabulário utilizado em tweets bêbados. Por outro lado, a semântica distribucional visa aprender as relações contextuais (sintáticas e semânticas) entre os termos de acordo com seu uso, fornecendo uma profunda compreensão das informações contextuais contidas nos tweets através de representações distribuídas, mais especificamente *word embeddings*⁴. *Word embeddings* demonstrou-se poderosa em várias aplicações relacionadas a classificação de textos (por exemplo, Kim (2014), Badjatiya et al. (2017)). Desta forma, este trabalho aplica *word embeddings* para lidar com as idiosincrasias da linguagem empregada em tweets bêbados. Nesta dissertação, exploram-se os ganhos obtidos na classificação de textos bêbados por essas duas abordagens de enriquecimento contextual e investiga como combinar seus pontos fortes.

Nesta dissertação, são propostos e comparados dois métodos para fornecer informações contextuais a tweets bêbados: *Drunk2Symbol* e *Drunk2Vec*. *Drunk2Symbol* tem como objetivo identificar *features* externas que forneçam significado e contexto aos tweets, usando a compreensão da linguagem natural e a Web Semântica. *Drunk2Symbol* também extrai *features* comportamentais relacionadas ao abuso do álcool. *Drunk2Vec* implementa uma rede neural convolucional para aprender *word embeddings* específicas do domínio que capturam as relações sintáticas e semânticas usadas em tweets bêbados. O *Drunk2Vec* baseia-se no *Drunk2Symbol* para integrar os pontos fortes dos dois métodos de enriquecimento contextual.

Para avaliar as contribuições dos métodos propostos, foi definido um amplo protocolo experimental envolvendo diferentes algoritmos de classificação, incluindo um conjunto de classificadores e cinco bases de dados. Além de explorar a base de dados elaborada por Hossain et al. (2016), foram construídas e disponibilizadas duas bases de dados adicionais que combinam as técnicas implementadas em Cavazos-Rehg et al. (2015) e Hossain et al. (2016). As bases de dados abrangem diferentes comportamentos de textos bêbados. Os resultados apresentaram melhorias significativas de até 13,7 na revocação e 12,2 pontos percentuais (pp) na precisão. Os experimentos revelaram que o enriquecimento contextual externo melhora a revocação através da extração de *features* simbólicas que generalizam para conceitos os termos utilizados em tweets. A semântica distribuci-

⁴O termo *word embeddings* é amplamente utilizado na literatura, desta forma o mesmo não foi traduzido

onal, por outro lado, contribui principalmente para melhorar a precisão, caracterizando os termos de acordo com seu uso. O uso de um conjunto de classificadores permitiu combinar e equilibrar as vantagens de cada técnica de enriquecimento contextual.

As principais contribuições deste trabalho são sumarizadas da seguinte forma:

- a proposta de dois métodos para fornecer contexto aos tweets. Esses métodos extraem *features* dos tweets e são utilizados com algoritmos de aprendizado de máquina para classificar textos bêbados. Os classificadores testados, utilizando as *features* produzidas pelos métodos propostos, superaram o *baseline* (HOSSAIN et al., 2016) em todas as métricas de avaliação. Ambos os métodos são independentes do algoritmo de classificação e podem ser utilizados com uma ampla gama de classificadores;
- uma avaliação detalhada das contribuições das estratégias de enriquecimento contextual para a classificação de textos bêbados. Para este fim, foi definido um protocolo experimental abrangente que engloba três classificadores distintos, cinco base de dados que abordam diferentes comportamentos relacionados ao consumo de álcool e uma análise detalhada das contribuições de cada estratégia de enriquecimento contextual e da integração de ambas as estratégias;
- uma análise exploratória dos dados, apresentando o potencial de informações relacionadas ao consumo de álcool que podem ser extraídas das redes sociais, sem a utilização de dados privados. Para este fim, foi elaborado um conjunto de visualizações que exploram os termos mais utilizados em tweets bêbados, a caracterização demográfica dos usuários, um estudo dos estabelecimentos vinculados e uma análise que compara as emoções expressas em tweets sóbrios e bêbados;
- a disponibilização de duas bases de dados públicas para a realização de novos experimentos com a classificação de textos bêbados.

O trabalho está organizado da seguinte forma. O Capítulo 2 aborda a fundamentação teórica e os conceitos que embasam os métodos propostos nesta dissertação. No Capítulo 3, são descritos e comparados os trabalhos relacionados. O Capítulo 4 apresenta as bases de dados propostas e utilizadas, além da análise exploratória dos dados. No Capítulo 5, são especificados os métodos propostos para a identificação de textos bêbados. O Capítulo 6 apresenta os experimentos e os resultados obtidos. O Capítulo 7 conclui esta dissertação, resumindo as contribuições deste trabalho, apresentando as limitações e os trabalhos futuros. Por fim, o Apêndice A detalha os resultados do Experimento #4.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda os conceitos necessários para o embasamento dos métodos propostos neste trabalho. Inicialmente, são definidos os algoritmos de aprendizado de máquina que são utilizados durante os experimentos. Em seguida, são apresentadas as técnicas de aprendizado profundo que embasam o método *Drunk2Vec* proposto nesta dissertação. Na sequência, são descritas abordagens complementares de redução de dimensionalidade e o conjunto de classificadores utilizado para combinar os métodos propostos. Por fim, são descritas as métricas de avaliação adotadas para comparar com os trabalhos relacionados.

2.1 Algoritmos de aprendizado de máquina

A tarefa de classificação utilizando aprendizado de máquina pode ser definida como a construção de um modelo preditivo capaz de descobrir os relacionamentos existentes no conjunto de treinamento, possibilitando ao modelo prever corretamente a classe de uma nova instância do conjunto de testes (AGGARWAL; ZHAI, 2012). A descoberta dos padrões existentes na base de dados é realizada através de algoritmos de aprendizado supervisionado, que aprendem a fazer classificações com base em observações anteriores.

Algoritmos de aprendizado de máquina são amplamente utilizados para automatizar a classificação de textos, especialmente para grandes volumes de dados. A classificação de textos é uma importante subárea da mineração de opiniões que foca na categorização de documentos de acordo com seu conteúdo (KORDE; MAHENDER, 2012). Nesta tarefa de classificação, os textos geralmente são representados por meio de *Bag of Words* (BoW) ou N-grama. BoW representa textos através de um conjunto de palavras, desconsiderando a ordenação e a relação entre as palavras no texto. Por outro lado, N-grama representa textos através de uma sequência contígua de palavras, permitindo representar a coocorrência entre duas ou mais palavras no texto. O grau de coocorrência é definido por uma janela customizável de tamanho N .

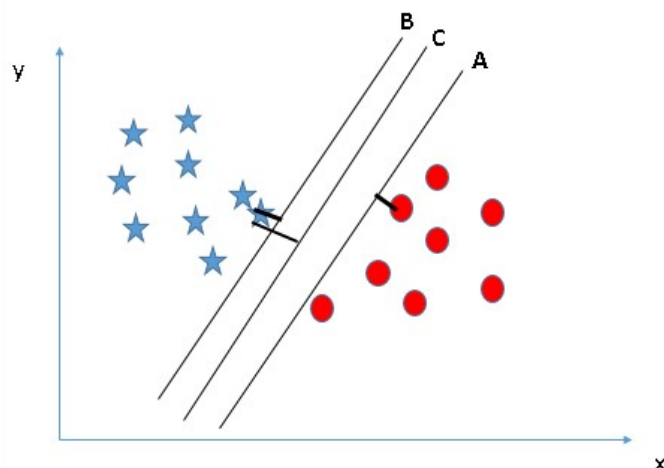
Existe uma ampla variedade de algoritmos de aprendizado de máquina disponíveis na literatura, como *Random Forest*, *Naïve Bayes*, *Support Vector Machines* (SVM) e *eXtreme Gradient Boosting* (XGBoost). A seguir, é descrito o funcionamento dos algoritmos SVM e XGBoost, pois ambos são aplicados neste trabalho. O algoritmo SVM foi implementado por todos os trabalhos relacionados na área, justificando sua implementa-

ção nesta dissertação. Por outro lado, o uso do XGBoost é motivado pelos bons resultados obtidos em competições de processamento de linguagem natural.

2.1.1 *Support Vector Machines*

Support Vector Machines (SVM) é um algoritmo de aprendizado supervisionado que constrói hiperplanos para mapear vetores de entrada em espaços não lineares de alta dimensionalidade. Cada dimensão deste espaço corresponde a um atributo do conjunto de dados de entrada. Após a construção dos hiperplanos, a classificação é realizada através da busca de um hiperplano ótimo, ou seja, que possui o melhor desempenho e apresenta a maior margem de separação para diferenciar as classes de entrada. A Figura 2.1 ilustra o funcionamento de um classificador criado com o algoritmo SVM, cujo melhor hiperplano é C. Esse hiperplano apresenta a maior margem de separação entre as classes.

Figura 2.1: SVM



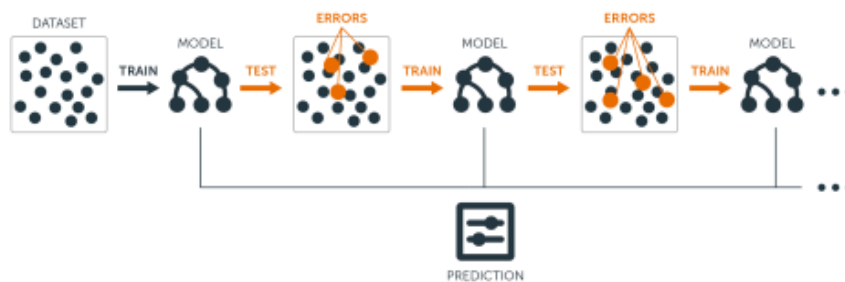
Fonte: <<https://pessoalex.wordpress.com/2019/04/10/algoritmo-svm-maquina-de-vetores-de-suporte-a-partir-de-exemplos-e-codigo-python-e-r/>>

O SVM foi aplicado com sucesso em uma variedade de problemas de PLN relacionados à classificação de dados obtidos a partir de redes sociais como, por exemplo, na identificação de crises (GHAFARIAN; YAZDI, 2020) e na detecção da depressão (PENG; HU; DANG, 2019).

2.1.2 *eXtreme Gradient Boosting*

Gradient Boosted Decision Trees (GBDT) é um algoritmo de classificação que combina árvores de decisão fracas para construir modelos preditivos eficazes, possuindo funcionamento similar ao *Random Forest*. O objetivo do classificador GBDT é minimizar a perda entre a classe prevista e a classe esperada (BOEHMKE; GREENWELL, 2019). Para este fim, iterativamente o GBDT adiciona novos modelos para melhorar o desempenho da classificação, de modo que a classificação final é atribuída através da análise de todas as árvores de decisão fracas construídas, conforme ilustrado na Figura 2.2. De forma complementar, *eXtreme Gradient Boosting*¹ (XGBoost) é uma biblioteca para otimizar o algoritmo GBDT, tornando-o escalável e altamente eficaz.

Figura 2.2: *Gradient Boosted Decision Trees*



Fonte: (BOEHMKE; GREENWELL, 2019)

O XGBoost figura entre os classificadores mais utilizados em soluções ganhadoras em competições da Kaggle² (CHEN; GUESTRIN, 2016), destacando-se principalmente em aplicações de Processamento de Linguagem Natural (PLN) como, por exemplo, na detecção de discurso de ódio em tweets (BADJATIYA et al., 2017) e na classificação hierárquica de textos (STEIN; JAQUES; VALIATI, 2019).

2.2 Aprendizado profundo

O aprendizado profundo é uma subárea do aprendizado de máquina que combina centenas de camadas com processamento não linear para extrair e transformar atributos de entrada em modelos preditivos (ZHANG; WANG; LIU, 2018). As camadas mais próximas à entrada de dados aprendem características simples, enquanto as camadas de alto

¹<https://xgboost.readthedocs.io/en/latest/>

²<https://www.kaggle.com/competitions>

nível aprendem características mais complexas derivadas das camadas anteriores.

As camadas utilizam modelos denominados redes neurais para ajustar os pesos dos neurônios e aprender as características vinculadas aos dados de entrada (CHOLLET; ALLAIRE, 2018). Para classificar textos, os modelos de redes neurais mais utilizados são: Redes Neurais Densas (DNN), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN). A seguir, é descrito o funcionamento das redes neurais CNNs e das *word embeddings*, pois ambas são implementadas nesta dissertação. O uso combinado de CNNs com *word embeddings* é motivado pelos bons resultados obtidos em tarefas tradicionais de PLN, incluindo análise de sentimentos e a classificação de textos curtos (KIM, 2014; ZHANG; WANG; LIU, 2018; AGARWAL et al., 2018; WANG et al., 2016).

2.2.1 Redes Neurais Convolucionais

Redes neurais convolucionais foram originalmente projetadas para classificar imagens e utilizadas na computação visual. Recentemente, as mesmas foram adaptadas para analisar textos. CNNs são especialmente utilizadas em tarefas de classificação cujo objetivo é encontrar padrões locais invariantes, ou seja, padrões que podem aparecer em qualquer lugar no conjunto de dados de entrada (GOLDBERG, 2016). Por exemplo, na análise de sentimentos de textos, uma sequência de palavras pode determinar o sentimento expresso, não importando a localização da sequência dentro do texto.

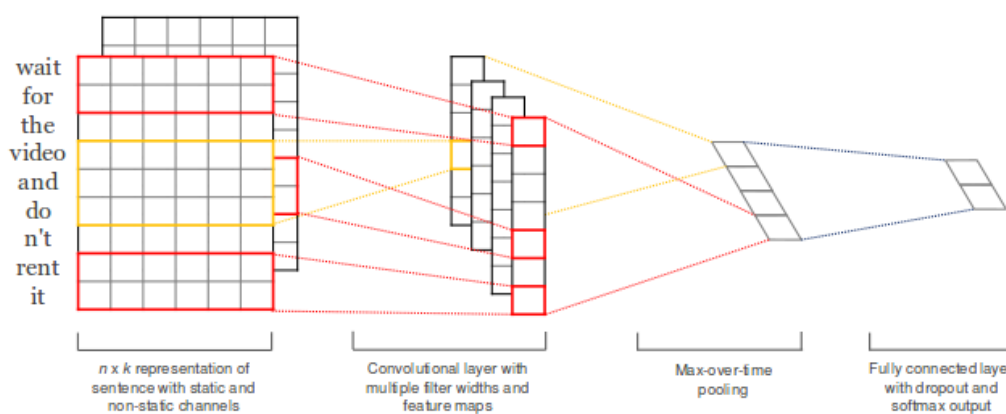
CNNs baseiam seu modelo de aprendizado em operações convolucionais e *pooling*, conforme ilustrado na Figura 2.3. A seguir, são descritas as quatro camadas principais em uma arquitetura típica de CNN para a classificação de textos:

- **input:** a entrada é uma representação vetorial dos *tokens* que constituem o conjunto de dados. Geralmente, utiliza-se *word embeddings* ou codificação *one-hot* para criar essa representação vetorial;
- **operação convolucional:** essa operação usa janelas de tamanhos predefinidos para extrair fragmentos da entrada. Em seguida, é aplicada a mesma transformação em todos os fragmentos, produzindo mapas de recursos (GOLDBERG, 2016). O objetivo dessa operação é reconhecer padrões locais em uma sequência (CHOLLET; ALLAIRE, 2018);
- **pooling:** o *pooling* combina mapas de *features* resultantes da operação convolucional em um vetor de dimensão única, reduzindo a amostra agressivamente. A

redução na amostra identifica as *features* mais importantes da sentença. Geralmente, o *pooling* representa o valor máximo ou médio observado em cada mapa de *features* (GOLDBERG, 2016);

- **camada totalmente conectada:** por último, é utilizada uma camada totalmente conectada responsável pela regularização e normalização. A regularização é aplicada para evitar o *overfitting*. Geralmente, essa operação faz uso de uma camada de *dropout*. *Dropout* elimina aleatoriamente alguns neurônios ocultos da rede neural, tornando o modelo mais generalizável. A normalização é composta por uma função de ativação que exporta as probabilidades correspondentes a(s) classe(s) alvo. Utiliza-se *softmax* para múltiplas probabilidades e *sigmoid* para a classificação binária.

Figura 2.3: Exemplo de arquitetura da CNN para PLN



Fonte: (KIM, 2014)

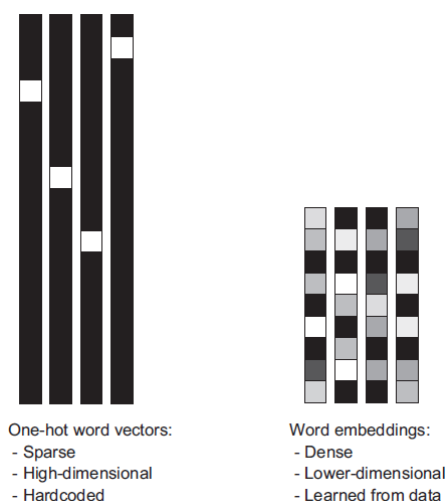
2.2.2 Representação textual para o aprendizado profundo

No aprendizado profundo, os documentos de entrada são representados usando codificação *one-hot* ou *word embeddings*. A codificação *one-hot* converte textos em uma matriz binária, onde as linhas representam as palavras e as colunas denotam os documentos. O valor inteiro um (1) indica que uma palavra ocorre em um documento, enquanto as demais posições da matriz são preenchidas com zero (0). O número de colunas da matriz é igual ao número de palavras presentes no vocabulário. Desta forma, *one-hot* representa os textos de forma esparsa e desconsidera a ordenação e o relacionamento entre as palavras no texto.

Enquanto a codificação *one-hot* representa dados em uma matriz esparsa de alta dimensionalidade, *word embeddings* é uma representação vetorial densa de baixa dimensionalidade aprendida a partir dos dados. A Figura 2.4 ilustra as estruturas da codificação *one-hot* e das *word embeddings*. A estrutura da esquerda exemplifica, em preto, a esparsidade da codificação *one-hot*, enquanto os diferentes tons de cinza da estrutura da direita representam a baixa dimensionalidade das *word embeddings*.

Word embeddings representam as relações semânticas e sintáticas, entre as palavras e seu contexto, aprendidas automaticamente a partir do conjunto de dados de entrada. Desta forma, palavras que frequentemente ocorrem em contextos similares são representadas próximas umas das outras na representação vetorial resultante. *Word embeddings* fornecem poder de generalização a um baixo custo computacional (GOLDBERG, 2017).

Figura 2.4: Codificação *One-hot* vs *Word embeddings*

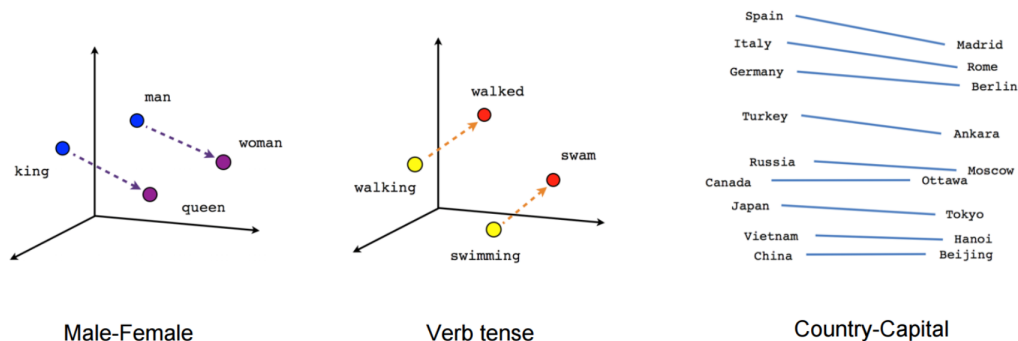


Fonte: (CHOLLET; ALLAIRE, 2018)

A Figura 2.5 exemplifica uma aplicação real das *word embeddings*. No exemplo, os vetores representam os relacionamentos entre gênero, tempo verbal e relacionamento semântico entre um país e sua capital. Por exemplo, adicionando o vetor ‘*man*’ ao vetor ‘*queen*’, identifica-se o vetor ‘*king*’. Assim, *word embeddings* podem ser empregadas para medir a similaridade entre palavras ou documentos (ZHANG; HE, 2018; GOODFELLOW et al., 2016).

Existem três abordagens geralmente utilizadas para aplicar *word embeddings*, conforme sumarizado a seguir:

- **utilizar algoritmos de propósito geral:** existem algoritmos projetados especificamente para gerar *word embeddings*. Os algoritmos mais populares são *Word2Vec*

Figura 2.5: Relacionamento entre *word embeddings*

Fonte: (MIKOLOV et al., 2013)

e *Global Vectors for Word Representation* (GloVe). O algoritmo *Word2Vec* propõe duas arquiteturas de rede neural para calcular *word embeddings*: *Continuous Bag of Words* (CBOW) e *Skip-gram*, conforme ilustrado na Figura 2.6. A primeira prevê uma palavra alvo a partir das palavras de contexto. Por exemplo na Figura 2.7, as palavras ‘cute’ e ‘jumps’ são usadas para prever o termo ‘cat’ quando o contexto é igual a um (1). Por outro lado, a arquitetura *Skip-gram* tenta prever as palavras de contexto a partir de um token alvo (MIKOLOV et al., 2013). *Skip-gram* é a arquitetura mais utilizada pela capacidade de compreender diferentes contextos, lidar com grandes conjuntos de dados e detectar palavras infrequentes. Outro método popular e poderoso para gerar *word embeddings* é GloVe, que se baseia na fatoração de matrizes de coocorrência entre as palavras em um determinado corpus (PENNINGTON; SOCHER; MANNING, 2014);

- ***word embeddings* pré-treinadas:** nesse modelo é necessário carregar um conjunto de *word embeddings* pré-treinadas. Existem repositórios abertos (FastText³, GloVe⁴ e Word2Vec⁵) que fornecem vetores já treinados em um grande volume de dados, para que possam ser aplicados em outros conjuntos de dados. No entanto, em bases de dados de domínio específico essa abordagem enfrenta problemas relacionados às palavras que estão fora do vocabulário (OOV), ou seja, quando os termos do conjunto de treinamento não fazem parte do vocabulário genérico usado para gerar as *word embeddings*. *Word embeddings* pré-treinadas podem ser usadas de forma estática ou dinâmica. Na primeira opção, é possível congelar a camada que representa os vetores de *word embeddings*, evitando que seus pesos sejam atualizados

³<https://fasttext.cc/>

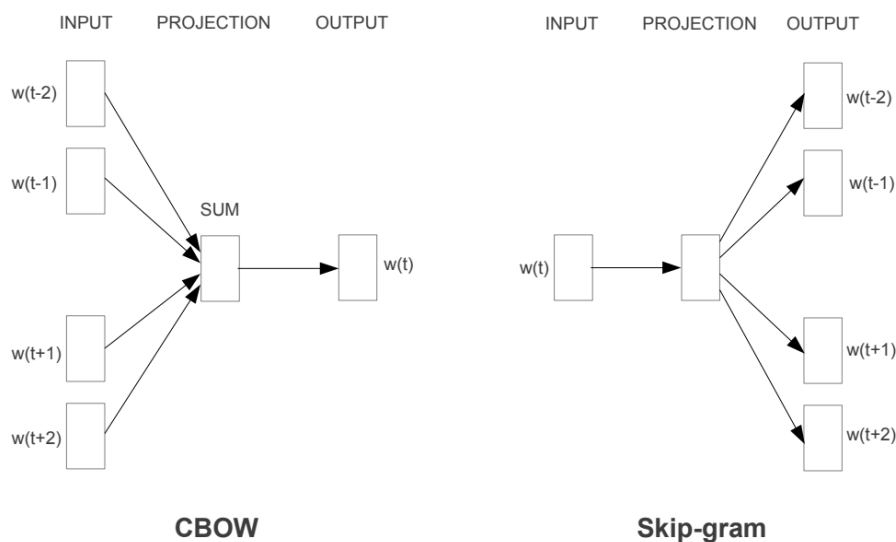
⁴<https://nlp.stanford.edu/projects/glove/>

⁵<https://code.google.com/archive/p/word2vec/>

durante o treinamento. Na opção dinâmica, os vetores de palavras são inicializados com *word embeddings* pré-treinadas e a atualização é propagada pela rede neural (CHOLLET; ALLAIRE, 2018);

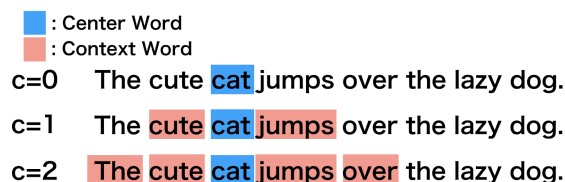
- **aprender *word embeddings* junto com a rede neural principal:** todos os vetores de palavras são inicializados aleatoriamente. Em seguida, iterativamente os vetores são atualizados da mesma maneira em que a rede neural aprende seus pesos, ou seja, pelo algoritmo de *backpropagation* (CHOLLET; ALLAIRE, 2018). Redes neurais CNN e *Long Short-Term Memory* (LSTM) são as mais implementadas para este fim. Essa abordagem fornece *word embeddings* específicas do domínio.

Figura 2.6: CBOW ou *Skip-Gram*



Fonte: (MIKOLOV et al., 2013)

Figura 2.7: Janela de contexto



Fonte: <<https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html>>

Word embeddings também foram implementadas para representar *features* de entrada e classificadas através de algoritmos tradicionais de aprendizado de máquina (por exemplo, SVM, XGBoost), produzindo resultados superiores ao estado da arte em (BENTON; ARORA; DREDZE, 2016; BADJATIYA et al., 2017; GHAFARIAN; YAZDI,

2020). Nesse caso, o texto é representado através de uma função de agregação que combina todas as palavras que compõem o texto, geralmente o vetor médio (BADJATIYA et al., 2017).

Nesta dissertação, aplica-se uma rede neural CNN para gerar *word embeddings* específicas para o domínio, uma vez que *word embeddings* de propósito geral enfrentam problemas relacionados as palavras que estão fora do vocabulário (OOV). Também foi utilizado o vetor médio das *word embeddings* para compor uma representação unidimensional dos tweets, visto que o vetor médio apresentou resultados superiores a utilização do *Doc2Vec* e *word embeddings* sem pré-processamento em experimentos preliminares. A Seção 5.3 exemplifica a metodologia para gerar *word embeddings* e calcular o vetor médio dos tweets.

2.3 Redução de dimensionalidade

A redução de dimensionalidade consiste no processo de diminuição do número de *features* em um determinado conjunto de dados, removendo *features* irrelevantes, redundantes e ruídos. Os algoritmos para este fim são classificados em seleção ou extração de *features*.

2.3.1 Algoritmos de seleção de *features*

Algoritmos de seleção de *features* identificam automaticamente o subconjunto das *features* com maior relevância dentro de um conjunto de dados. Esse processo melhora a eficiência dos classificadores de texto, pois auxilia na construção de modelos mais generalizáveis com menor custo computacional. De modo geral, estes algoritmos calculam a entropia ou baseiam-se em métodos estatísticos para selecionar apenas as *features* relevantes que estão relacionadas com o resultado, removendo *features* redundantes/irrelevantes do conjunto de dados.

Os algoritmos de seleção de *features* mais populares são *Correlation-based Feature Selection* (CFS) e *Information Gain* (InfoGain) (TANG; ALELYANI; LIU, 2014). O *InfoGain* calcula a entropia entre a presença de uma *feature* e a classe de destino, selecionando apenas as *features* relevantes (SEO, 2018). Nesta dissertação, foi utilizado o algoritmo *InfoGain* para selecionar as *features* oriundas do enriquecimento contextual ex-

terno que estão relacionadas com o resultado, evitando a introdução de *features* irrelevantes que podem prejudicar o desempenho dos classificadores. Também foram realizados experimentos com o algoritmo CFS, mas os resultados foram inferiores.

2.3.2 Algoritmos de extração de *features*

O objetivo dos algoritmos de extração de *features* também é reduzir a dimensionalidade do conjunto de dados de entrada. No entanto, estes algoritmos projetam as *features* de entrada em um novo espaço com baixa dimensionalidade (TANG; ALELYANI; LIU, 2014). Dessa forma, não é possível relacionar as *features* de entrada com as *features* projetadas. A Análise de Componentes Principais (PCA) e a Análise Discriminante Linear (LDA) são exemplos de técnicas de extração de *features*. O PCA aplica operações matemáticas para transformar as *features* de entrada em uma representação linear não correlacionada. Nesta dissertação, o algoritmo PCA foi implementado para projetar as *features* obtidas através do conhecimento externo em uma representação densa, evitando a criação de matrizes esparsas.

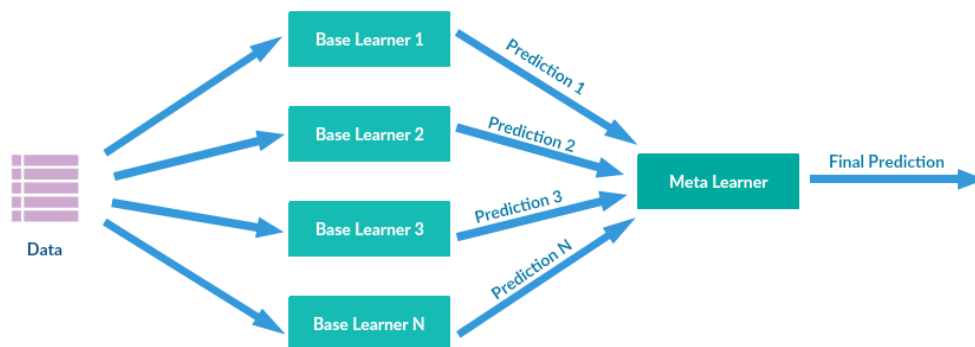
2.4 Conjunto de classificadores

Uma alternativa para melhorar o desempenho no aprendizado de máquina é combinar vários modelos preditivos (WITTEN et al., 2016). A combinação de modelos preditivos é denominada conjunto de classificadores. Essa combinação maximiza o desempenho de classificação, pois é capaz de combinar as previsões dos modelos iniciais, explorar as características principais de cada modelo e evitar *overfitting*.

O empilhamento é uma das formas mais utilizadas para criar um conjunto de classificadores. Neste método, ilustrado na Figura 2.8, são construídos vários modelos preditivos que exportam uma probabilidade correspondente ao grau de confiança para a classe alvo. As previsões iniciais são combinadas por um classificador supervisor, que utiliza funções de agregação ou algoritmos de aprendizado de máquina.

Neste trabalho, cada classificador é treinado individualmente e exporta uma probabilidade entre 0 e 1, representando o grau de confiança em que um determinado tweet é bêbado. Em seguida, é aplicada a técnica de empilhamento para combinar os classificadores e a média simples como função de agregação. Em outras palavras, a classificação

Figura 2.8: Conjunto de classificadores com empilhamento



Fonte: <<https://www.kdnuggets.com/2019/01/ensemble-learning-5-main-approaches.html/>>

final é calculada através da média das probabilidades classificadas individualmente. Essa combinação resulta em uma distribuição de probabilidade que reflete cada classificador. A função de agregação média foi selecionada por apresentar resultados superiores a utilização de algoritmos de aprendizado de máquina, por exemplo *Naïve Bayes*, para combinar as probabilidades individuais.

2.5 Métricas de avaliação

Através das métricas de avaliação é possível analisar o desempenho dos métodos de classificação, comparar diferentes abordagens e identificar o método mais eficaz. Foram aplicadas métricas de avaliação comumente utilizadas na recuperação de informações e por trabalhos relacionados na identificação de textos bêbados. A seguir, são sumarizadas as métricas utilizadas e as fórmulas de cálculo:

- Precisão (Equação 2.1): proporção de tweets bêbados classificados corretamente considerando todos os tweets classificados como bêbados
- Revocação (Equação 2.2): proporção de tweets bêbados classificados corretamente considerando todos os tweets cuja classe esperada é bêbado
- Medida F_1 (Equação 2.3): é obtida através da média harmônica entre a precisão e a revocação.

$$P = \frac{TP}{TP + FP} \quad (2.1)$$

$$R = \frac{TP}{TP + FN} \quad (2.2)$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (2.3)$$

Figura 2.9: Matriz de confusão

		Classe prevista	
		Positivo	Negativo
Classe real	Positivo	TP	FN
	Negativo	FP	TN

Fonte: Elaborado pelo autor

Para calcular as métricas descritas acima, é necessário identificar os indicadores TP, TN, FP e FN através da matriz de confusão (Figura 2.9). A seguir, é descrito o propósito de cada indicador:

- TP (Verdadeiro Positivo): classificação correta na classe positiva;
- TN (Verdadeiro Negativo): classificação correta na classe negativa;
- FP (Falso Positivo): classificação incorreta na classe positiva;
- FN (Falso Negativo): classificação incorreta na classe negativa.

Este trabalho usa a média da medida F_1 , precisão e revocação para avaliar os métodos propostos. Os indicadores são calculados através da média da classe positiva, ou seja, micro média.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta o contexto no qual essa dissertação está inserida. Inicialmente, apresenta-se a motivação e os trabalhos relacionados que exploram dados provenientes de redes sociais para extrair informações relevantes a partir de textos escritos por usuários possivelmente alcoolizados. Em seguida, são apresentados os trabalhos focados na classificação automática de tweets bêbados, destacando os principais desafios e as técnicas empregadas. Na parte final deste capítulo, são apresentados os trabalhos que utilizam técnicas de enriquecimento contextual para melhorar a classificação de textos curtos.

3.1 Motivação para identificar e classificar textos bêbados

Coletar informações sobre o consumo de álcool por meio de estudos especializados pode ser caro e demorado. Por outro lado, as redes sociais fornecem informações de alto valor com menor custo e em tempo real (CULOTTA, 2013). Trabalhos relacionados confirmaram a importância dos dados provenientes de redes sociais demonstrando que dados oficiais sobre o consumo de álcool são fortemente correlacionados a dados referentes ao consumo de álcool extraídos por meio de tweets (CULOTTA, 2013; WEST et al., 2012; CURTIS et al., 2018). Jovens que usam redes sociais têm maior probabilidade de usar tabaco, álcool e maconha (JOHNSON; SHAPIRO; TOURANGEAU, 2011). Além disso, a exposição repetida a postagens envolvendo drogas nas redes sociais pode incentivar outras pessoas, pois esse comportamento pode ser percebido como normal (WEST et al., 2012). Jernigan et al. (2017) constataram uma correlação entre o nível de exposição ao marketing e o nível de consumo de álcool entre os jovens, pois jovens expostos repetidamente ao marketing vinculado ao consumo de álcool têm maior probabilidade de começar a ingerir bebidas alcoólicas ou aumentar o consumo.

As redes sociais fornecem informações relevantes para monitorar e entender problemas de saúde pública, inclusive o consumo excessivo de álcool (CURTIS et al., 2018). As intervenções públicas que utilizam redes sociais podem ajudar a reduzir o consumo nocivo de álcool (KNOX et al., 2019). Aproximadamente 500 milhões de tweets são postados diariamente, sendo que 37% dos usuários registrados têm entre 18 e 29 anos. Portanto, o Twitter é uma fonte rica de dados em tempo real a partir dos quais padrões sobre o consumo de álcool podem ser extraídos (WEST et al., 2012). Enviar mensagens

de texto sob influência do álcool é conhecido popularmente como *texto bêbado*. Nas redes sociais, denomina-se *tweet bêbado* qualquer tweet escrito sob influência do álcool (JOSHI et al., 2015).

A importância em identificar mensagens de textos bêbados é ampliada ao considerarmos sua aplicabilidade na extração de informações relevantes de usuários alcoolizados sem depender de processos manuais ou dados privados (FILHO; CARVALHO; PAPPAS, 2014). West et al. (2012), Cavazos-Rehg et al. (2015) e Maity et al. (2018) buscam identificar o papel das redes sociais na disseminação e na prevenção do consumo de álcool através da caracterização demográfica de usuários que postam textos bêbados. West et al. (2012) e Maity et al. (2018) utilizam técnicas de visualização para identificar características comuns a textos bêbados (por exemplo, período, evento, comportamento e aspectos linguísticos). Cavazos-Rehg et al. (2015) analisou sentimentos e tópicos abordados em tweets bêbados. Roller, Thomas e Schmeier (2018) exploraram a influência da Copa do Mundo de 2018 no consumo de bebidas alcoólicas, analisando os países e as partidas de futebol com maior consumo de álcool. Todos os trabalhos mencionados realizaram suas análises em tweets classificados por palavras-chave como, por exemplo, *drunk* e sinônimos.

O uso de palavras-chave para classificar tweets é limitado, pois pode falhar ao considerar expressões com negação, sendo suscetível à ambiguidade léxica (HASAN; AGU; RUNDENSTEINER, 2014; MYSLÍN et al., 2013). Também é frequente a adoção de gírias (*tipsy*, *zosted*, etc) em tweets bêbados e nem todos os tweets contêm termos explícitos que se referem ao consumo de álcool. Da mesma forma, nem todos os tweets que contêm palavras-chave específicas (por exemplo, *drunk*) se referem ao álcool (por exemplo, *drunk in love*). Portanto, é necessário investir na classificação automática de textos bêbados, permitindo assim uma análise rápida e confiável de um grande conjunto de dados a baixo custo.

3.1.1 Identificação automática de textos bêbados

A identificação automática de textos bêbados consiste no processo de classificar um documento em {bêbado, sóbrio} de acordo com o seu conteúdo. Jauch, Jaehne e Suenemann (2013), Aphinyanaphongs et al. (2014), Joshi et al. (2015), Hossain et al. (2016) e Maity et al. (2018) abordaram este processo de classificação utilizando aprendizado de máquina. Todos esses trabalhos se baseiam em um conjunto de dados de treinamento

para obter resultados de qualidade e variam nos recursos explorados e nos algoritmos de classificação.

Embora o termo tweet bêbado refira-se a tweets escritos sob a influência de álcool (JOSHI et al., 2015), não existe um conjunto de dados oficial contendo textos escritos por pessoas alcoolizadas. A base de dados *Alcohol Language Corpus* (ALC) contém discursos em que pessoas sóbrias e bêbadas realizam atividades pré-determinadas (por exemplo, descrever figuras e contar números) e é limitado ao idioma alemão (SCHIEL; HEINRICH; BARFÜSSER, 2012). Para superar essa limitação, Aphinyanaphongs et al. (2014), Joshi et al. (2015), Hossain et al. (2016) e Maity et al. (2018) construíram seus conjuntos de dados com base na presença de palavras-chave específicas (*drunk* e sinônimos) e realizaram ações para anotar os dados. Enquanto Aphinyanaphongs et al. (2014) definiram um protocolo e anotaram manualmente seus tweets, Hossain et al. (2016) utilizaram o serviço de *crowdsourcing* Amazon Mechanical Turk¹ (AMT). Em contrapartida, Joshi et al. (2015) e Maity et al. (2018) usaram palavras-chave para anotar automaticamente os tweets.

Jauch, Jaehne e Suendermann (2013), Aphinyanaphongs et al. (2014), Joshi et al. (2015), Hossain et al. (2016) e Maity et al. (2018) usaram técnicas tradicionais de extração de *features* para a classificação de textos, como pré-processamento (por exemplo, tokenização, remoção de *stopwords*, normalização de menções/URLs), extração de n-gramas e análise de sentimentos. Joshi et al. (2015) também analisaram a presença de caracteres repetidos, erros ortográficos e características de estilo (conectores de discurso, tamanho, etc). *Hashtags* foram incorporadas e analisadas por todos trabalhos relacionados que utilizam tweets, destacando sua importância em análises aplicadas em redes sociais. Todos os trabalhos mencionados anteriormente implementaram o algoritmo SVM e testaram diferentes algoritmos como, por exemplo, *Random Forest*, *Generalized Linear Model*, *Logistic regression* e *Naïve Bayes*.

3.2 Enriquecimento contextual

Bons resultados na classificação de textos dependem do tamanho e da qualidade dos textos dos quais as *features* são extraídas (LI et al., 2016). No entanto, tweets são curtos e esparsos, possuem ruídos, utilizam um vocabulário próprio e caracterizam-se pela falta de contexto. Abordagens para lidar com esses desafios são enriquecer o contexto dos

¹<http://www.mturk.com>

tweets com enriquecimento externo ou *word embeddings*.

3.2.1 Enriquecimento contextual externo

O enriquecimento contextual externo expande o conteúdo do tweet com recursos externos representativos que são relacionados à classe alvo. De modo geral, o enriquecimento externo utiliza bases de conhecimento externas (e.g., DBPedia, Wikipédia) e ferramentas de compreensão da linguagem natural (e.g., Open Calais, IBM Watson) para adicionar contexto a textos curtos. As principais melhorias estão relacionadas à expansão e generalização do vocabulário usando o conhecimento extraído de fontes externas de conhecimento e, portanto, são referidas neste trabalho como enriquecimento contextual externo.

Os trabalhos relacionados à classificação de textos bêbados não abordam o enriquecimento contextual externo. Desta forma, para exemplificar casos de sucesso do enriquecimento contextual externo, são apresentados trabalhos similares em outras áreas de classificação em tweets. Romero e Becker (2019), Schulz, Guckelsberger e Janssen (2017) utilizaram ferramentas de reconhecimento de entidades nomeadas e a DBPedia para melhorar a classificação de eventos a partir de tweets. Melhorias foram observadas na revocação. De modo similar, Mizzaro et al. (2014) baseou-se na Wikipédia para enriquecer seus tweets e na API do Google para capturar páginas web relacionadas, apresentando melhorias na categorização de tweets.

3.2.2 *Word embeddings*

Outra alternativa cada vez mais popular para adicionar contexto é a semântica distribucional, representada usando *word embeddings*. *Word embeddings* fornecem poder de generalização ao relacionar palavras com contexto sintático ou semântico similar em representações vetoriais de baixa dimensionalidade, capturando uma infinidade de relacionamentos identificados na linguagem utilizada.

Em combinação com o aprendizado profundo, a semântica distribucional foi implantada com sucesso em várias aplicações que classificam tweets, como na análise de sentimentos (KIM, 2014; HARB; BECKER, 2018), transtornos mentais (YATES; COHAN; GOHARIAN, 2017), epidemias (KHATUA; KHATUA; CAMBRIA, 2019),

consumo de drogas ilícitas (HU et al., 2018) e detecção de ódio (BADJATIYA et al., 2017). Alguns desses trabalhos observaram melhorias com o uso de *word embeddings* pré-treinadas (KIM, 2014; HARB; BECKER, 2018). Por outro lado, Hu et al. (2018) e Khatua, Khatua e Cambria (2019) aplicaram *Word2Vec* para gerar embeddings específicas do domínio, indicando que essa abordagem é melhor do que *word embeddings* pré-treinadas e genéricas para domínios altamente específicos, mesmo para corpus relativamente menores (KHATUA; KHATUA; CAMBRIA, 2019). A partir de um grande conjunto de dados, Yates, Cohan e Goharian (2017) geraram *word embeddings* específicas do domínio usando pesos inicializados aleatoriamente como parte de uma CNN para classificar a depressão. Badjatiya et al. (2017) propuseram uma rede neural LSTM para gerar *word embeddings* específicas do domínio que capturam as especificidades da linguagem abusiva, uma vez que os usuários empregam gírias, jargões e variações, dificultando a detecção automática dos sistemas de moderação.

3.3 Análise dos trabalhos relacionados

Esta seção apresenta uma análise comparativa dos trabalhos descritos neste capítulo. Para realizar esta comparação, foi definido um conjunto de características principais dos trabalhos relacionados na classificação de textos bêbados. A Tabela 3.1 sumariza as principais características e os diferenciais deste trabalho. Nessa tabela, os seguintes símbolos denotam que: (i) ✓ - o trabalho aborda essa característica; (ii) ✗ - o trabalho não aborda essa característica; (iii) N/A - a característica não é aplicável ao trabalho. Os trabalhos estão identificados na Tabela 3.1 pelo acrônimo das três primeiras letras do autor principal: Jauch, Jaehne e Suendermann (2013) é denominado como ‘JAU’, Aphinyanaphongs et al. (2014) é representado por ‘APH’, Joshi et al. (2015) é ‘JOS’, Hossain et al. (2016) possui como sigla ‘HOS’, Maity et al. (2018) é ‘MAI’. As características e os trabalhos relacionados são descritas a seguir.

A ‘*fonte de dados*’ para classificação é a primeira característica abordada. Devido à inexistência de uma base de dados oficial, Aphinyanaphongs et al. (2014), Joshi et al. (2015), Hossain et al. (2016) e Maity et al. (2018) construíram suas bases de dados utilizando tweets no idioma inglês. Jauch, Jaehne e Suendermann (2013) foi o único trabalho a utilizar a base de dados ALC que contém discursos em alemão de pessoas sob influência do álcool. Algumas características (‘*filtro por palavras-chave*’, ‘*processo de anotação*’, ‘*hashtags*’, ‘*emoticons*’) não se aplicam a Jauch, Jaehne e Suendermann

(2013), pois a base de dados ALC é disponibilizada de forma anotada e não possui dados de redes sociais.

A característica '*filtro por palavras-chave*' refere ao conjunto de palavras (*seeds*) utilizadas para pré-selecionar os tweets. Os trabalhos utilizaram diferentes conjuntos de palavras-chave para coletar os dados. Joshi et al. (2015) empregaram as *hashtags* *#drunk*, *#drank*, *#imdrunk*, *#imnotdrunk* e *#sober* para coletar os tweets. Hossain et al. (2016) utilizaram termos relacionados a intoxicação alcoólica (por exemplo, *drunk*, *alcohol*, *booze*, *vodka*, *hangover*, *wasted*). Maity et al. (2018) utilizaram sinônimos do termo *drunk* para selecionar os tweets (por exemplo, *Drunk*, *Wasted*, *Tipsy*, *Intoxicated*, *Hammered*, *Sauced*, *Buzzed*). Por outro lado, Aphinyanaphongs et al. (2014) não utilizaram palavras-chave como filtros, selecionando todos tweets georreferenciados, em Nova Iorque, próximos a virada do ano de 2012.

Após coletar os dados, é necessário identificar os tweets da classe positiva e negativa. Esse processo é descrito na característica '*processo de anotação*'. Hossain et al. (2016) utilizaram o serviço de *crowdsourcing* Amazon Mechanical Turk. Aphinyanaphongs et al. (2014) definiram um protocolo que analisa os seguintes itens para determinar se o tweet é bêbado: i) ato de consumir álcool; ii) intenção de beber; iii) localização de um bar ou loja de bebidas; iv) menção de uma marca específica; v) *hashtags* relacionadas ao álcool. Joshi et al. (2015) e Maity et al. (2018) usaram palavras-chave para anotar automaticamente os tweets. Portanto, Aphinyanaphongs et al. (2014) e Hossain et al. (2016) apresentaram as estratégias mais robustas para anotar os dados, visto que não se baseiam em palavras-chave.

Hashtags são frequentemente utilizadas em redes sociais, especialmente em tweets. Desta forma, todos os trabalhos aplicados a redes sociais fizeram uso das informações contidas nas '*hashtags*' para o processo de classificação.

Joshi et al. (2015), Hossain et al. (2016) e Maity et al. (2018) abordaram a '*análise de sentimentos*', visto que o consumo excessivo de álcool afeta na forma com que as emoções são expressas (BORRILL; ROSEN; SUMMERFIELD, 1987). A Seção 4.5.4 apresenta um comparativo entre as emoções expressas por usuários sóbrios e bêbados. Hossain et al. (2016) analisaram os sentimentos representados nos *emoticons*. Em contrapartida, Joshi et al. (2015) e Maity et al. (2018) fizeram uso de dicionários léxicos para analisar os sentimentos expressos nos tweets.

A característica '*morfologia da sentença*' refere-se à análise gramatical da sentença, identificando, por exemplo, adjetivos, substantivos e conjunções. Joshi et al. (2015)

foram os únicos a analisar a morfologia da sentença, extraindo o número de conjunções e a taxa entre adjetivos, substantivos e advérbios que constam nos tweets. Os resultados apresentados por Joshi et al. (2015) indicam que a análise de palavras (N-grama) é mais eficaz que a análise de características de estilo da sentença, entre elas a morfologia. Desta forma, não foram implementadas análises morfológicas neste trabalho.

Uma das premissas avaliadas é que usuários sob influência do álcool tendem a escrever tweets com erros de digitação e erros ortográficos, comparados na tabela com a característica ‘*verificação ortográfica*’. Desta forma, Joshi et al. (2015) utilizaram um validador ortográfico para verificar os tweets. Joshi et al. (2015) e Hossain et al. (2016) também analisaram a presença de alongamentos, visto que usuários costumam repetir uma letra da palavra para expressar sua opinião (GUPTA; JOSHI, 2017), comparados na tabela através do item ‘*caracteres repetidos*’.

Todos os trabalhos utilizaram o ‘*aprendizado de máquina*’ como estratégia de classificação e aplicaram o algoritmo SVM. Também se destaca a utilização dos algoritmos *Naïve Bayes* e *Logistic Regression*, que foram implementados por Jauch, Jaehne e Suendermann (2013), Aphinyanaphongs et al. (2014) e Maity et al. (2018). Todos os trabalhos também utilizam representações em nível de palavras (N-grama ou BoW) como ‘*engenharia de features*’. Joshi et al. (2015), Hossain et al. (2016) e Maity et al. (2018) utilizaram N-grama. Em contrapartida, Jauch, Jaehne e Suendermann (2013) utilizaram apenas BoW. Aphinyanaphongs et al. (2014) utilizaram ambas as representações (N-grama e BoW).

‘*Enriquecimento externo*’, ‘*seleção de features*’, ‘*word embeddings*’ e ‘*aprendizado profundo*’ não foram abordados pelos trabalhos apresentados. Desta forma, este trabalho é pioneiro da adoção destas técnicas para identificar textos bêbados.

3.4 Considerações finais

Neste capítulo foram apresentados os principais tópicos relacionados à classificação automática de textos bêbados e o potencial de informações relacionadas ao abuso do álcool que podem ser extraídas. Também foram descritas duas estratégias de enriquecimento contextual e suas aplicações em tarefas relacionadas a classificação de textos em redes sociais. Por fim, foi elaborado um comparativo entre os trabalhos relacionados.

Uma das principais dificuldades relatadas pelos trabalhos analisados é a falta de uma base de dados oficial relacionada aos textos bêbados. Para contornar essa limita-

Tabela 3.1: Comparativo entre os trabalhos relacionados que realizam classificação automática de textos bêbados

Característica	JAU	APH	JOS	HOS	MAI	Este trabalho
Fonte de dados	ALC	tweets	tweets	tweets	tweets	tweets
Filtro por palavras-chave	N/A	×	hashtags indicando consumo de álcool	termos relacionados ao consumo de álcool	sinônimos de drunk	termos relacionados ao consumo de álcool
Processo de anotação	N/A	manual	manual	manual (Mturk)	manual	manual (Mturk)
Hashtags	N/A	✓	✓	✓	✓	✓
Análise de sentimento	×	×	✓	✓	✓	✓
Morfologia da sentença	×	×	✓	×	×	×
Verificação ortográfica	×	×	✓	×	×	✓
Caracteres repetidos	×	×	✓	✓	×	✓
Aprendizado de máquina	✓	✓	✓	✓	✓	✓
Engenharia de features	BoW	N-grama e Bow	N-grama	N-grama	N-grama	N-grama
Enriquecimento externo	×	×	×	×	×	✓
Seleção de features	×	×	×	×	×	✓
Word Embeddings	×	×	×	×	×	✓
Aprendizado profundo	×	×	×	×	×	✓

ção, este trabalho explora tweets da língua inglesa, assim como a maioria dos trabalhos relacionados. O Twitter foi utilizado devido a sua popularidade e disponibilidade de dados. As bases de dados fornecidas por Hossain et al. (2016) e Cavazos-Rehg et al. (2015) são exploradas neste trabalho. Adicionalmente, este trabalho constrói uma base de dados composta por tweets relacionados ao consumo de álcool para demonstrar a importância das informações extraídas das redes sociais e para verificar a eficácia dos métodos propostos. Maiores informações sobre as bases de dados são descritas no Capítulo 4.

Todos os trabalhos utilizaram o aprendizado supervisionado como estratégia de classificação e representações em nível de palavras. Entre os trabalhos mencionados, Hossain et al. (2016) alcançou o melhor desempenho (medida F_1 de 88,29%) usando o algoritmo SVM sobre *features* textuais, particularmente termos frequentes, *hashtags* e sentimentos dos *emojicons*. No entanto, nenhum dos trabalhos explorou informações contextuais que podem lidar com os ruídos e a esparcialidade dos tweets, nem com técnicas de aprendizado profundo.

Desta forma, os esforços deste trabalho concentraram-se na exploração de estratégias que fornecem contexto aos tweets. A primeira estratégia, denominada enriquecimento contextual externo, provê contexto e generalizações aos tweets. Sua aplicação foi motivada pelo fato que tweets não contêm palavras suficientes para fornecer o contexto necessário para classificá-los. Em especial, destaca-se o trabalho da Romero e Becker

(2019), cuja as técnicas de enriquecimento externo serviram como inspiração para este trabalho. A segunda estratégia adota *word embeddings* e aprendizado profundo. As técnicas de aprendizado profundo foram baseadas em Badjatiya et al. (2017), visto que é uma solução interessante para abordar as idiossincrasias do vocabulário utilizado em textos bêbados, um tipo de contexto que não é tratado adequadamente pelo enriquecimento contextual externo. Este trabalho é pioneiro na adoção de ambas as estratégias de enriquecimento contextual para classificar textos bêbados.

4 BASES DE DADOS - TWEETS RELACIONADOS AO CONSUMO DE ÁLCOOL

Este capítulo é dedicado à apresentação das bases de dados que contêm tweets relacionados ao consumo de álcool e que são utilizadas para avaliar os experimentos dos métodos propostos neste trabalho. A primeira é uma base de dados utilizada em outros trabalhos da literatura atual e as outras duas são propostas e construídas especificamente para este trabalho. Inicialmente, é fornecida uma visão geral das bases de dados. Em seguida, cada uma delas é detalhadamente apresentada. Por fim, é apresentada a análise exploratória dos dados coletados para a terceira base de dados.

4.1 Visão geral

Para avaliar os métodos propostos nesta dissertação, são utilizadas três bases de dados compostas por tweets que abordam diferentes comportamentos vinculados ao consumo de álcool no Twitter. A primeira base de dados, denominada *DS1-drunk*, tem como objetivo avaliar os métodos em tweets que mencionam bebidas alcoólicas, o consumo de álcool e o consumo de álcool enquanto utiliza o Twitter. A finalidade do *DS2-keywords* é verificar o desempenho dos métodos em uma amostra aleatória de tweets classificados através de palavras-chave. A última base de dados, denominada *DS3-drinking-ext*, é empregada em conjunto com técnicas de visualizações para ilustrar a importância das informações extraídas das redes sociais e para avaliar o desempenho dos métodos em uma base de dados de médio tamanho.

A Tabela 4.1 fornece uma visão geral das bases de dados, apresentando as regras usadas para anotar os tweets e o tamanho de cada base de dados. As bases de dados *DS1-drunk* e *DS3-drinking-ext* utilizaram anotadores e possuem 3.994 e 4.456 tweets, respectivamente. *DS1-drunk* foi produzido por (HOSSAIN et al., 2016), enquanto *DS3-drinking-ext* foi construído especificamente para este trabalho. Por outro lado, *DS2-keywords* é uma extensão de Cavazos-Rehg et al. (2015) e possui 11.792 tweets classificados através de palavras-chave. Todas as bases de dados são compostas por tweets escritos na língua inglesa. Essa escolha foi motivada pela existência de mais ferramentas para processamento de linguagem natural, bases de conhecimento e serviços de *crowdsourcing* para o idioma inglês.

Tabela 4.1: Resumo das bases de dados

Característica	Base de dados		
	DS1-drunk	DS2-keywords	DS3-drinking-ext
Coleta (palavras-chave)	Bebidas alcoólicas e consumo de álcool	<i>Drunk, beer, alcohol, liquor, vodka e hangover</i>	Consumo de álcool
Anotação	Amazon Mechanical Turk	Palavras-chave	Amazon Mechanical Turk
Tamanho	3.994 tweets	11.792 tweets	4.456 tweets
Objetivo do trabalho	Identificar tweets bêbados	Analisar tópicos e sentimentos	Identificar tweets bêbados
Trabalhos relacionados	Hossain et al. (2016)	Cavazos-Rehg et al. (2015)	Este trabalho

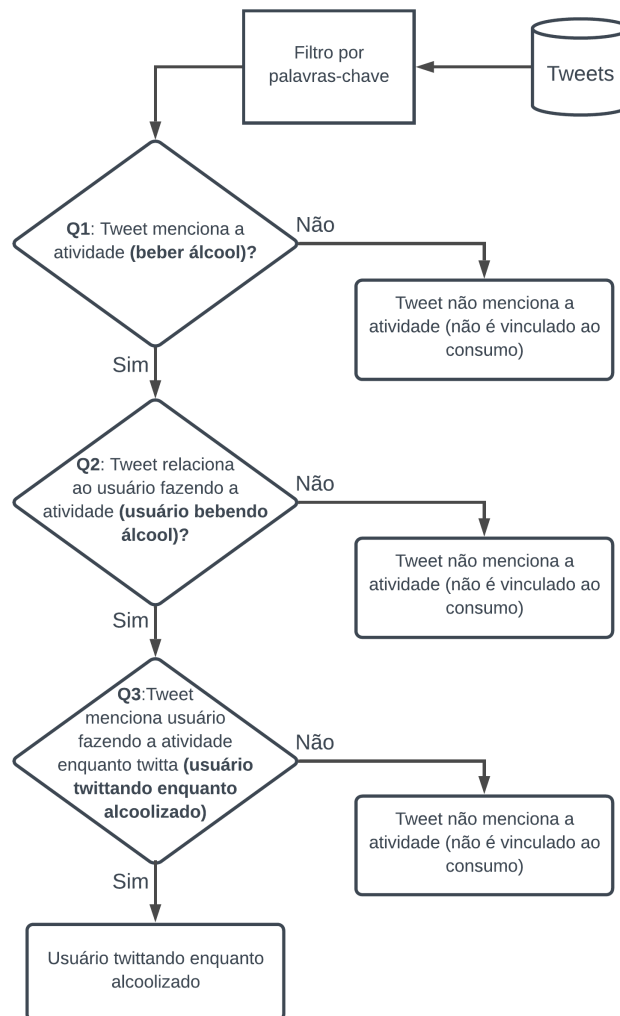
4.2 Base de dados *DS1-drunk*

O objetivo desta base de dados é avaliar o desempenho dos métodos propostos na classificação de três diferentes temáticas relacionadas ao consumo de álcool. As temáticas envolvem a análise de menções a bebidas alcoólicas, o consumo de álcool e se o usuário está consumindo álcool e utilizando as redes sociais. *DS1-drunk* é um subconjunto do disponibilizado por Hossain et al. (2016) que é composto de tweets para identificar o consumo de álcool. Utilizou-se um subconjunto dos dados originais, visto que nem todos os tweets estavam disponíveis para download. Assim, foi possível recuperar 3.996 tweets dos 5.559 tweets do conjunto de dados original. É importante destacar que apenas foi realizado o processo de download dos tweets existentes, ou seja, não foi necessário nenhum processo adicional de coleta ou anotação.

A Figura 4.1 apresenta o processo de coleta e anotação dos dados desta base de dados. Inicialmente, foram coletados tweets que contêm menções ao consumo de álcool. Em seguida, cada avaliador respondeu três perguntas relacionadas a cada tweet. Esses processos são detalhados nas próximas subseções.

4.2.1 Coleta de dados

O processo de coleta dos tweets baseou-se na busca de tweets georreferenciados, mais especificamente no estado de Nova Iorque, que continham palavras-chave pré-determinadas. A lista de palavras-chave é baseada em referências a bebidas alcoólicas (por exemplo, *beer* e *vodka*), estados alcoólicos (por exemplo, *hangover* e *buzzed*), entre outras.

Figura 4.1: Fluxograma para criação do *DSI-drunk*

Fonte: Adaptado de Hossain et al. (2016)

4.2.2 Anotação de dados

Após coletar os tweets, os mesmos foram anotados através do *Amazon Mechanical Turk*¹ (AMT) de acordo com as três perguntas descritas a seguir e conforme ilustrado na Figura 4.1:

- Q1: tweet menciona a atividade (**beber álcool**)
- Q2: tweet relaciona ao usuário fazendo a atividade (**usuário bebendo álcool**)
- Q3: tweet menciona o usuário fazendo a atividade enquanto tuíta (**usuário tuitando enquanto alcoolizado**)

Cada tweet foi anotado por três avaliadores, de modo que a classificação final foi

¹<http://www.mturk.com>

atribuída através da maioria de votos. *DS1-drunk-mention* representa o conjunto de tweets relacionados a pergunta *Q1*. Os tweets vinculados a *Q2* constam na base de dados *DS1-drunk-drinking*. *DS1-drunk-drink-now* simboliza os tweets vinculados a *Q3*. O número total de registros, detalhados pelas classes positiva e negativa, assim como as regras de anotação são sumarizados na Tabela 4.2.

Tabela 4.2: Características do *DS1-drunk*

Base de dados	#Tweets	Regra	#Tweets bêbados	#Tweets sóbrios
DS1-drunk-mention	3.994	Menciona o consumo de álcool	2.236	1.758
DS1-drunk-drinking	1.777	Usuário consumindo álcool	1.374	403
DS1-drunk-drink-now	1.054	Usuário consumindo álcool enquanto tuíta	651	403

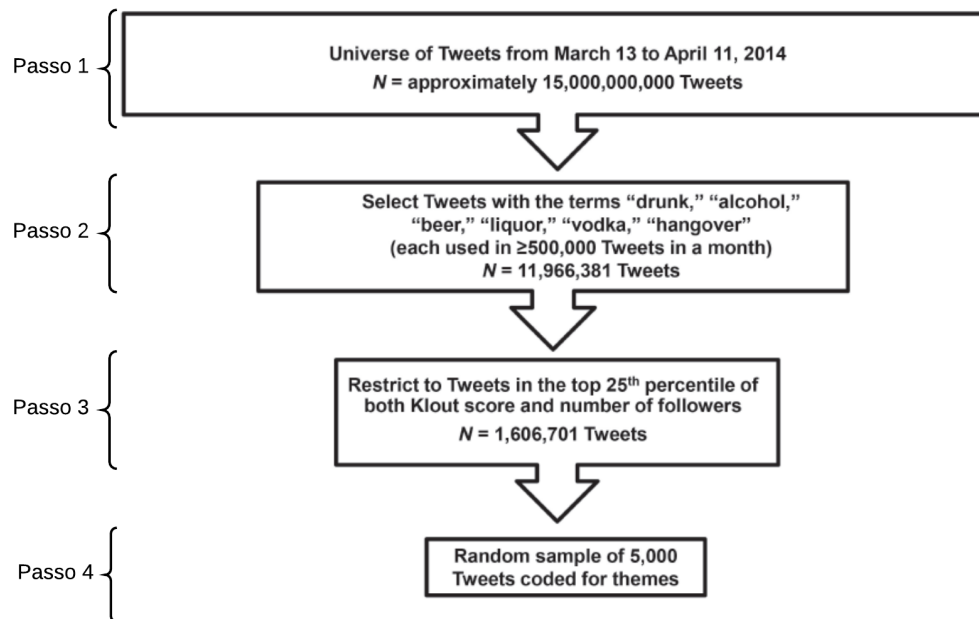
4.2.3 Dicionário de dados

Os itens a seguir descrevem a estrutura desta base de dados. Ressalta-se que Hos-sain et al. (2016) apenas forneceram o identificador dos tweets e a classificação consolidada (*Q1*, *Q2*, *Q3*). As demais informações foram extraídas diretamente do Twitter para esta dissertação.

- *tweetid*: identificador único do tweet
- *tweet_text*: conteúdo do post
- *link*: url de acesso ao tweet
- *retweetid*: identificador do tweet relacionado
- *mentions*: menções e *hashtags*
- *user_name*: nome do usuário no Twitter
- *date*: data de inclusão do post
- *q1*: resultado consolidado da primeira pergunta
- *q2*: resultado consolidado da segunda pergunta
- *q3*: resultado consolidado da terceira pergunta
- *geo*: coordenadas geográficas

4.3 Base de dados *DS2-keywords*

O objetivo desta base de dados é identificar tweets que contêm menções ao consumo de álcool ou bebidas alcoólicas. Esta base de dados é composta por uma amostra

Figura 4.2: Fluxograma para coleta do *DS2-keywords*

Fonte: Adaptado de Cavazos-Rehg et al. (2015)

de tweets classificados através de palavras-chave. Mais especificamente, *DS2-keywords* é uma extensão dos dados fornecidos por Cavazos-Rehg et al. (2015) para analisar tópicos e sentimentos em tweets bêbados. Porém, Cavazos-Rehg et al. (2015) apenas disponibiliza tweets classificados como bêbados. Desta forma, foi necessário estender o trabalho relacionado incorporando tweets classificados como sóbrios. Outra característica importante de *DS2-keywords* é o processo automático de anotação, ou seja, a classificação foi realizada de acordo com a presença de palavras-chave. A seguir, é apresentada a metodologia de coleta de dados e de anotação.

4.3.1 Coleta de dados

A Figura 4.2 ilustra o fluxograma que descreve o processo de coleta de dados aplicado Cavazos-Rehg et al. (2015). No primeiro passo, foi coletada uma amostra aleatória de tweets. No passo seguinte, foram filtrados tweets que contêm um dos seguintes termos: “*drunk*”, “*#drunk*”, “*alcohol*”, “*#alcohol*”, “*beer*”, “*#beer*”, “*liquor*”, “*#liquor*”, “*vodka*”, “*#vodka*”, “*hangover*” ou “*#hangover*”. No terceiro passo, foram selecionados os tweets mais influentes. Por fim, foi extraída uma amostra de 5.000 tweets. Este conjunto de dados pode ser relacionado em significado ao *DS1-drunk-mention* (beber álcool).

Para criar um conjunto correspondente da classe negativa, identificou-se os usuários que postaram os tweets da classe positiva. Em seguida, foram coletados os demais tweets publicados que não continham as palavras-chave listadas na seção anterior. Este processo de coleta resultou em 6.792 tweets sóbrios. O limite máximo foi de 5 tweets por usuário. A motivação em usar os mesmos usuários é para avaliar se há alterações na maneira como usuários sob influência do álcool se expressam.

4.3.2 Anotação de dados e pré-processamento

Os tweets fornecidos por Cavazos-Rehg et al. (2015) foram classificados como positivos (bêbado). Por outro lado, os demais tweets foram classificados como negativos (sóbrio). Por fim, foram removidos todos os termos usados durante a coleta para evitar classificá-los com base nos mesmos padrões usados para obter os dados. Os tweets são disponibilizados em um repositório público².

4.3.3 Dicionário de dados

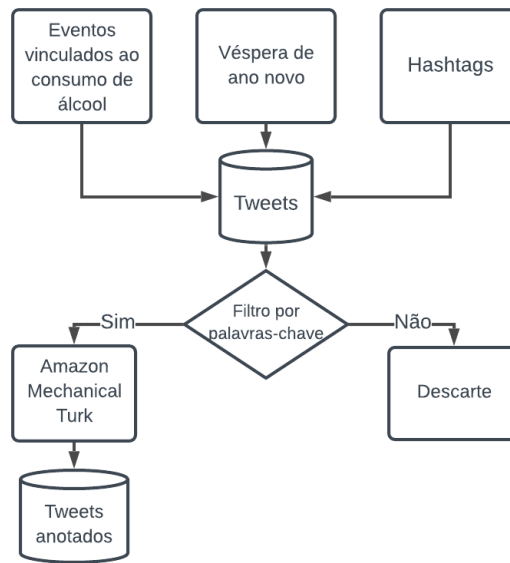
A estrutura deste conjunto de dados inclui os campos elencados a seguir:

- *followers*: número de seguidores
- *link*: url de acesso ao tweet
- *retweetid*: identificador do tweet relacionado
- *tweet_text*: conteúdo do post
- *tweetid*: identificador único do tweet
- *mentions*: menções e *hashtags*

4.4 Base de dados *DS3-drinking-ext*

O objetivo geral do *DS3-drinking-ext* é construir uma base de dados de tamanho médio que esteja vinculada a usuários que consomem álcool, possibilitando a realização de análises exploratórias e a classificação automática dos dados. Para este fim, foram coletados tweets relacionados ao consumo de álcool. Em seguida, os tweets foram classi-

²https://gist.github.com/MarcosGrzeca/40bb6b38dc88b6ad76ad01e29af79b38#file-ds2_sobertweets

Figura 4.3: Fluxograma para criação do *DS3-drinking-ext*

Fonte: Elaborado pelo autor

ficados via o serviço de *crowdsourcing* Amazon Mechanical Turk, conforme apresentado na Figura 4.3. Através do *DS3-drinking-ext* é possível averiguar a eficácia dos classificadores em bases de dados de diferentes tamanhos. A seguir, são fornecidos detalhes relacionados aos processos de coleta, anotação dos dados e o dicionário de dados.

4.4.1 Coleta de dados

Foram coletados tweets da língua inglesa que contêm termos relacionados a eventos vinculados ao consumo de álcool (*Saint Patrick's Day*, *Mardi Gras*, *Oktoberfest*), tweets contendo *hashtags* vinculadas ao álcool (*#drunk*, *#imdrunk*, *#drank*) e tweets publicados na véspera de Ano Novo (31/12 a 01/01) entre 2017 e 2018, inclusive. A lista completa de palavras-chave é disponibilizada em um repositório público³. Uma vez que a API do Twitter não permite a coleta de postagens antigas, utilizou-se a ferramenta *GetOldTweets-python*⁴ para obter os tweets retroativos.

Em seguida, foram selecionados tweets que contêm menções ao consumo de álcool. Para filtrar os tweets, foi criada uma lista de palavras-chave que combina os filtros propostas por West et al. (2012) e Cavazos-Rehg et al. (2015). A lista é composta pelos termos: *'Drunk'*, *'Wasted'*, *'Topsy'*, *'Intoxicated'*, *'Hammered'*, *'Sauced'*, *'Buzzed'*,

³<https://gist.github.com/MarcosGrzeca/e69cc31539114600b73eb8e12ec791c0>

⁴<https://github.com/Jefferson-Henrique/GetOldTweets-python>

Figura 4.4: Ferramenta para anotar *DS3-drinking-ext*


Instructions

Judge if the tweet is related to alcohol consumption


Examples to help with answers:

Tweet example	Response
So lets just stay in the moment, smoke some weed, drink some wine	Yes
Kendall Day before the wedding, wedding reception with open bar, then after party downtown. Holy shit tomorrow is going to kick ass.	Yes
@CollinDSchuck Its alcohol and emotions I hope Montreal takes it just as seriously. Who can forget all the rally towel related injuries haha	No
@_illegitOut_: @NaiShekirah I graduate this year turn up*aye me too, Turnup . Class 2014	No

Tweet:



Killa Coop
@Christina_2_U



Bree is drunk. I am drunk

♥ 3:16 AM - May 1, 2014 · Syracuse, NY ⓘ

👤 See Killa Coop's other Tweets >

Tweet is about the tweeter drinking alcohol?

Yes
 No
 Not sure

Fonte: Elaborado pelo autor

'Trashed', 'Under the influence', 'Blacked out', 'Gassed', 'Zooted', 'Juiced', 'Designated driver', 'Shit-faced', 'Plastered', 'Inebriated', 'Cock eyed', 'Cockeyed', 'Shwasted', 'Zonked', #drunk', '#drank' e '#imdrunk'.

4.4.2 Anotação dos dados

O processo de anotação é o ponto mais crítico na criação de bases de dados, pois exige a colaboração de avaliadores humanos e é suscetível ao entendimento de cada avaliador. O serviço *Amazon Mechanical Turk* foi utilizado para anotar os tweets deste conjunto de dados. Os avaliadores responderam a mesma pergunta usada para rotular *DS1-drunk-drinking* (HOSSAIN et al., 2016), ou seja, *'Tweet is about the tweeter drinking alcohol?'* As opções disponíveis de resposta foram *'Yes'*, *'No'* e *'Not sure'*, conforme apresentado na Figura 4.4. Cada tweet foi avaliado por três anotadores mestres, ou seja, anotadores que mostraram excelente desempenho nas tarefas em que trabalharam. Por fim, os tweets foram classificados de acordo com a maioria das respostas. Por exemplo, se pelo menos dois avaliadores responderem *'Yes'*, o tweet foi classificado como texto bêbado.

Entre os 2.963 tweets selecionados, 1.774 foram anotados como tweets bêbados⁵. Para medir a concordância entre os anotadores, utilizou-se o algoritmo *Fleiss' Kappa*. *Fleiss' Kappa* calcula um índice estatístico que representa a confiabilidade e a qualidade da base de dados (FALOTICO; QUATTO, 2015). A taxa *Fleiss' Kappa* é de 0,292 nesta base de dados, indicando concordância razoável (LANDIS; KOCH, 1977). Os níveis de concordância do índice *Kappa* são descritos no Apêndice A. Este conjunto de dados é um complemento para *DS1-drunk-drinking*, então combinou-se os 1.933 tweets contidos em *DS1-drunk-drinking*, para criar o conjunto de dados *DS3-drinking-ext*.

4.4.3 Dicionário de dados

A seguir, é descrita a estrutura da base de dados *DS3-drinking-ext*:

- *tweetid*: identificador único do tweet
- *tweet_text*: conteúdo do post
- *link*: url de acesso ao tweet
- *retweetid*: identificador do tweet relacionado
- *mentions*: menções e *hashtags*
- *user_name*: nome do usuário no Twitter
- *date*: data de inclusão do post
- *a1, a2, a3*: respostas atribuídas pelos avaliadores
- *drunk*: classificação
- *geo*: coordenadas geográficas

4.5 Análise exploratória dos dados

O objetivo da análise exploratória é apresentar o potencial de informações que podem ser extraídas das redes sociais com a associação entre textos e perfis de usuários, sem a necessidade de dados privados. Dessa forma, é demonstrado que dados públicos extraídos de tweets bêbados, classificados automaticamente, podem identificar fatores que proporcionam uma melhor compreensão dos perfis de usuários que consomem álcool e utilizam o Twitter.

⁵<https://gist.github.com/MarcosGrzeca/40bb6b38dc88b6ad76ad01e29af79b38#file-ds3-tweets-labeled>

A base de dados *DS3-drinking-ext* foi utilizada para desenvolver as seguintes análises relacionadas ao consumo de álcool:

- nuvem de palavras contendo os termos mais utilizados em tweets bêbados;
- pirâmide demográfica dos usuários;
- locais mais frequentados pelo público consumidor de álcool;
- análise de sentimentos.

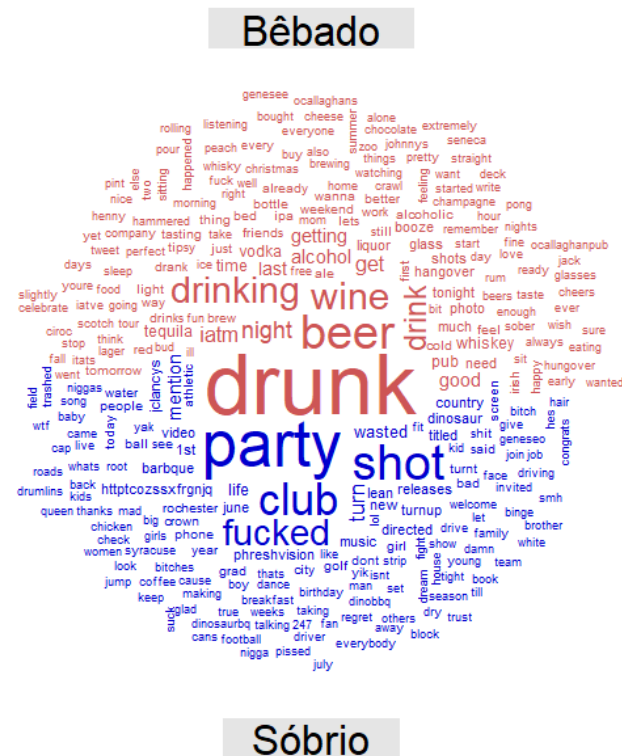
4.5.1 Termos mais utilizados por usuários que estão sob efeito do álcool

O objetivo desta análise é verificar os termos mais utilizados por usuários que estão sob influência do álcool. Para isso, foi construída uma nuvem de palavras (Figura 4.5) contendo os 300 termos mais usados por usuários sóbrios (azuis) e bêbados (vermelhos). Observe que os termos *party*, *shot* e *club* estão entre os mais frequentes no grupo sóbrio, pois a base de dados foi construída através da filtragem prévia de palavras-chave que mencionam o consumo de álcool (*drunk*, *beer*, *party*, etc...). O termo *pissed* é utilizado em tweets sóbrios e bêbados, pois sua definição depende do contexto e pode significar ‘chateado’ ou ‘bêbado’. No grupo relacionado ao consumo de álcool, existem referências explícitas a bebidas alcoólicas (*wine*, *beer*, *tequila* e *vodka*) entre as palavras mais comuns.

Também foram analisadas as *features* conceituais extraídas através do enriquecimento contextual. De forma geral, *features* conceituais⁶ são extraídas através de ferramentas de processamento de linguagem natural e descrevem conceitos atrelados aos tweets. A Figura 4.6 exibe uma nuvem de palavras contendo as 200 *features* conceituais mais frequentes. As *features* conceituais mais empregadas em tweets bêbados são referências a bebidas alcoólicas (*foodanddrink*, *beverages*, *alcoholicbeverages*, *cocktail-sandbeer* e *wine*). Por outro lado, atividades relacionadas ao lazer (*artandentertainment*, *moviesandtv*, *shopping* e *lawgovtandpolitics*) destacam-se no grupo de tweets sóbrios. As *features* conceituais *familyandparenting* e *programminglanguages* estão relacionadas aos tweets bêbados, porém as mesmas não são diretamente relacionadas ao consumo de álcool. Para verificar essas *features*, foram analisados os tweets que as contêm. Conclui-se que a *feature programminglanguages* é relacionada aos tweets que contêm a palavra *typing*, enquanto a *feature familyandparenting* associa tweets que contêm relacionamentos

⁶As *features* conceituais e as ferramentas utilizadas são descritas na Seção 5.2.3

Figura 4.5: Nuvem de palavras para tweets sóbrios e bêbados



Fonte: Elaborado pelo autor

familiares, sendo comum o consumo de álcool em festas familiares de final de ano.

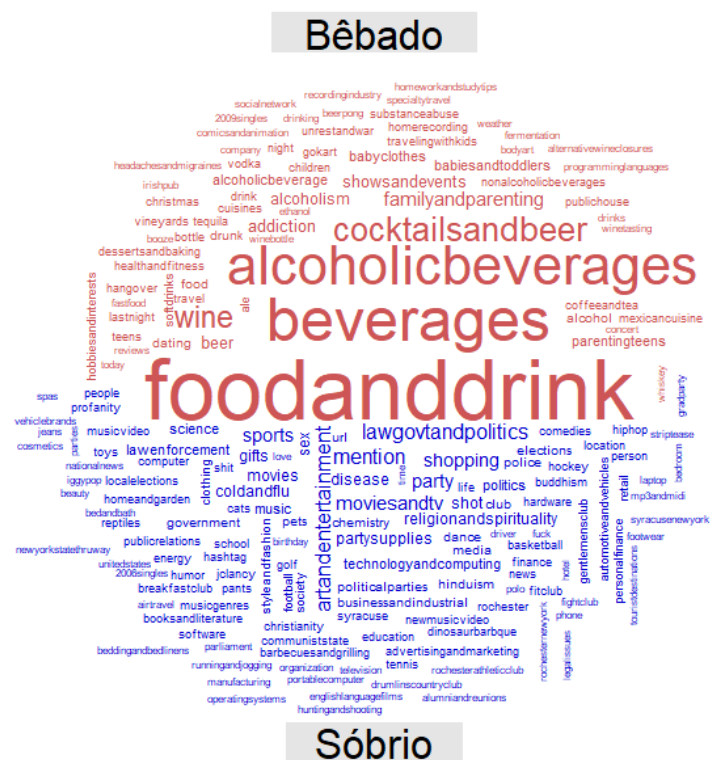
4.5.2 Pirâmide demográfica

O intuito desta visualização é identificar o público mais afetado pelo consumo de álcool. Para extrair essas informações, os usuários do Twitter foram agrupados por gênero e faixa etária em uma pirâmide demográfica. No entanto, o Twitter não solicita dados pessoais dos usuários. Para contornar essa limitação, foram aplicadas ferramentas que inferem dados pessoais analisando apenas informações públicas disponibilizadas pelos usuários. A idade dos usuários foi obtida através da verificação da foto do perfil com o Face++⁷. Em seguida, as idades foram discretizadas em cinco faixas etárias, conforme An e Weber (2016). Com o objetivo de identificar o gênero dos usuários, combinou-se as ferramentas Face++ e Genderizer⁸. O Face++ identifica o gênero através da análise da imagem do perfil, enquanto o Genderizer identifica o gênero considerando apenas o nome

⁷<https://www.faceplusplus.com>

⁸<https://api.genderize.io>

Figura 4.6: Nuvem de palavras comparando as *features* conceituais



Fonte: Elaborado pelo autor

do usuário. Nos casos em que o gênero retornado pelo Genderizer divergiu do Face++, uma análise manual do perfil foi realizada verificando o usuário, URL externa, nome e as fotos.

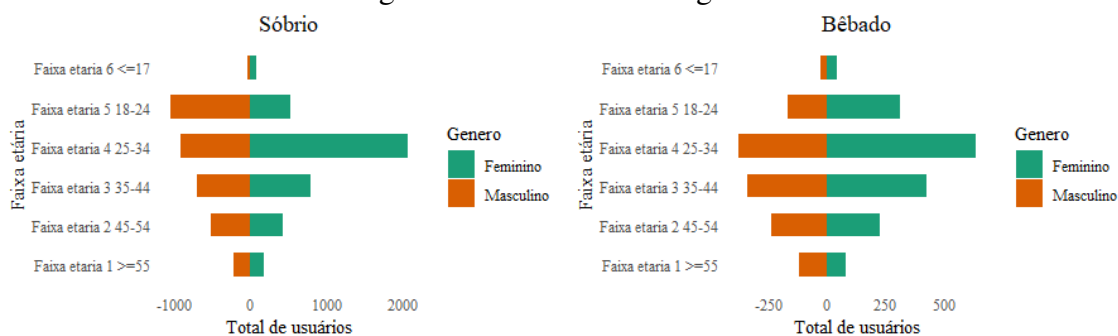
A Figura 4.7 exibe duas pirâmides demográficas. A pirâmide da esquerda representa os usuários sóbrios, enquanto a pirâmide da direita sumariza os usuários sob influência do álcool. Nesta base de dados, há mais mulheres e adultos. Considerando esse viés, não se observou diferenças significativas em relação ao gênero ou faixa etária que diferencia tweets sóbrios e bêbados.

4.5.3 Estabelecimentos vinculados ao consumo de álcool

O objetivo deste estudo é verificar se o consumo excessivo de álcool se concentra em um grupo de estabelecimentos. Desta forma, agrupou-se os estabelecimentos em categorias (por exemplo, bares, pubs, restaurantes) com o uso da API do Facebook Places⁹. O Facebook Places foi utilizado para buscar os estabelecimentos localizados em um raio

⁹<https://developers.facebook.com/docs/places>

Figura 4.7: Pirâmide demográfica



Fonte: Elaborado pelo autor

de 100 metros da localização original do tweet. Em seguida, extraiu-se a categoria do estabelecimento mais próximo ao tweet. Essa análise foi possível, pois aproximadamente 40% dos tweets nesta base de dados são georreferenciados.

Estabelecimentos classificados como restaurantes, pubs e bares têm um número maior de tweets relacionados do que os demais (Figura 4.8). Ao agrupar as categorias de localização em grupos (Figure 4.9), nota-se que estabelecimentos relacionados ao grupo *Food & Beverage* são os locais mais próximos dos usuários neste conjunto de dados. O grupo *Food & Beverage* engloba as categorias *American Restaurants*, *Pubs*, *Bar & Grill*, *Bar*, *Sports Bar*, entre outras. Também é importante ressaltar que o grupo *Education*, em especial a categoria *The College & University*, é o segundo grupo mais frequente, indicando uma alta taxa de consumo de álcool nas universidades. Os tweets vinculados a categoria *The College & University* estão principalmente concentrados em quatro universidades locais: *The College at Brockport State University of New York*, *SUNY Cortland*, *Syracuse University* e *Monroe Community College*.

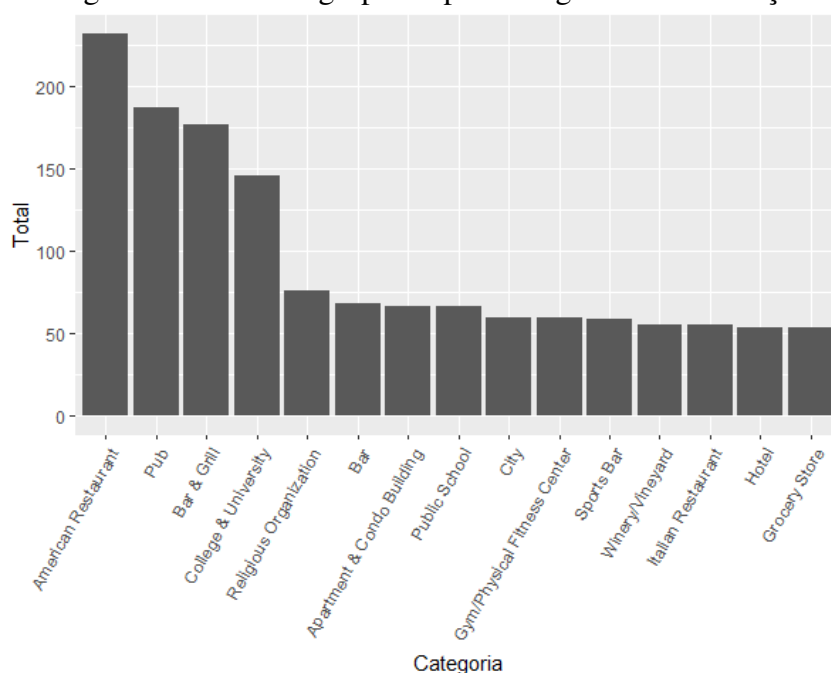
4.5.4 Análise de sentimentos

A última análise busca investigar se existem diferenças na forma como os sentimentos são expressos por usuários que estão sob influência do álcool (JOSHI et al., 2015; BORRILL; ROSEN; SUMMERFIELD, 1987). Para este fim, foi aplicado o dicionário de sentimentos NCR¹⁰ para extrair as emoções expressas nos tweets.

Os resultados são exibidos na Figura 4.10. É possível observar que os tweets bêbados estão relacionados a felicidade (21%), enquanto, nos tweets sóbrios, essa emoção

¹⁰<http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>

Figura 4.8: Tweets agrupados pela categoria da localização



Fonte: Elaborado pelo autor

é expressa apenas em 13% dos tweets. Tweets sóbrios expressam emoções variadas, como tristeza (13%), raiva (15%) e medo (14%). A proporção de tweets que contém essas emoções reduz significativamente em tweets bêbados (9%, 11% e 9%, respectivamente).

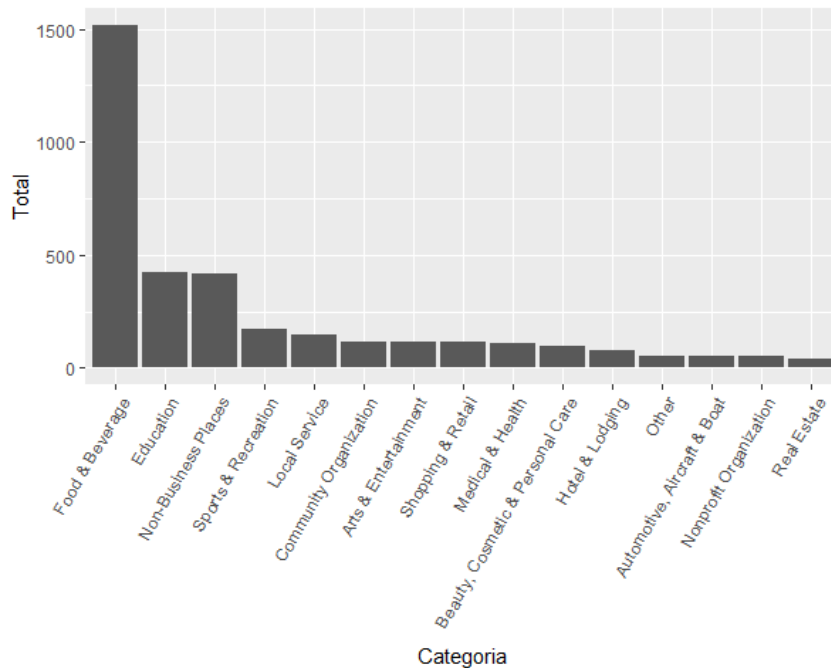
4.6 Considerações finais

Este capítulo descreveu as três bases de dados para a realização dos experimentos que avaliam os métodos propostos. Cabe enfatizar que as três bases de dados podem ser aplicadas em análises para identificar o papel das redes sociais na disseminação e na prevenção do consumo de álcool e para classificar automaticamente tweets bêbados, permitindo avaliar e comparar o desempenho de diferentes métodos.

As bases de dados possuem como limitação o idioma inglês, mas as técnicas de coleta e anotação podem ser reaproveitadas para outros idiomas. Devido às políticas de privacidade do Twitter, apenas é permitido disponibilizar publicamente o identificador dos tweets, impedindo a publicação de seu conteúdo. Por fim, devido às políticas de acesso do *Amazon Mechanical Turk*, não foi possível anotar um grande volume de tweets para *DS3-drinking-ext*, uma vez que nossas contas foram suspensas, impedindo a anotação de novos tweets.

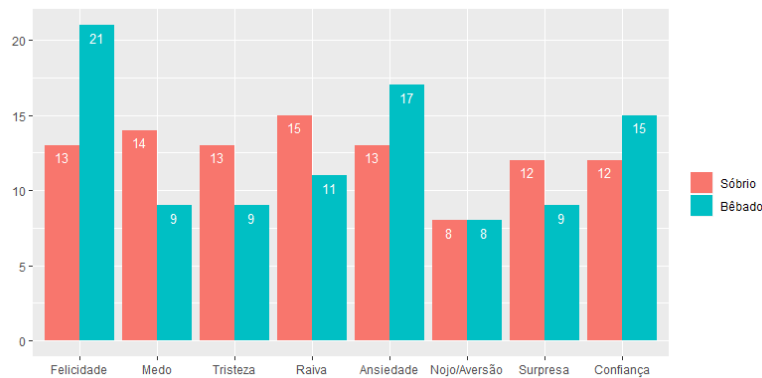
Para ilustrar a aplicabilidade da classificação de mensagens de texto alcoolizada,

Figura 4.9: Tweets organizados pelo grupo da categoria de localização



Fonte: Elaborado pelo autor

Figura 4.10: Sentimentos expressos nos tweets



Fonte: Elaborado pelo autor

foi desenvolvida a análise exploratória com os termos frequentes, *features* conceituais, dados demográficos, localidades e a representatividade das emoções. A análise foi desenvolvida em uma base de dados limitada (*DS3-drinking-ext*), mas confirmou a relação entre estudantes de faculdades/universidades e bebidas alcoólicas, bem como a mudança de emoções que podem explicar a motivação para ingerir bebidas alcoólicas (por exemplo, redução da tristeza e raiva, aumento da alegria) e a incidentes (por exemplo, redução do medo). No entanto, essas suposições precisam ser confirmadas por um estudo mais abrangente com uma base de dados maior. A análise exploratória pode ser estendida a outros problemas sociais ou para visualizar dados coletados a partir das redes sociais.

5 DRUNK2SYMBOL E DRUNK2VEC: MÉTODOS DE ENRIQUECIMENTO CONTEXTUAL PARA A CLASSIFICAÇÃO DE TEXTOS BÊBADOS

Este capítulo apresenta as principais contribuições deste trabalho: a proposta de dois métodos que fornecem informações contextuais aos tweets para melhorar a classificação de textos bêbados. Inicialmente é apresentado o primeiro método, cuja principal característica é a exploração do enriquecimento externo contextual. Em seguida, são fornecidos detalhes relacionados ao segundo método que se caracteriza pela utilização da semântica distribucional.

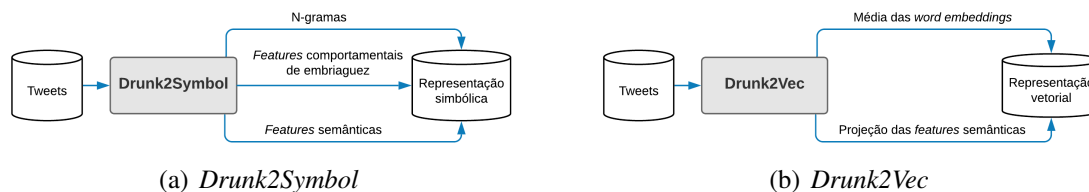
5.1 Visão geral

Os métodos propostos são denominados *Drunk2Symbol* e *Drunk2Vec* e produzem *features* explorando estratégias distintas de enriquecimento contextual. O *Drunk2Symbol* tem como objetivo principal prover significado e generalizações aos tweets através da exploração de bases externas de conhecimento que fornecem contexto. O *Drunk2Vec* combina a riqueza semântica das *word embeddings* com *features* contextuais externas, assumindo que essas duas técnicas de enriquecimento são complementares: enquanto o *Drunk2Symbol* expande e generaliza o vocabulário para lidar com a esparcialidade, *word embeddings* refletem as especificidades do vocabulário utilizado em tweets bêbados. O *Drunk2Vec* implementa uma rede neural CNN para aprender *word embeddings* que representam as relações sintáticas e semânticas dos termos usados em tweets bêbados, criando representações vetoriais de tweets que podem ser classificadas a partir de diferentes algoritmos.

A Figura 5.1 destaca as diferentes *features* produzidas por cada método. *Drunk2Symbol* utiliza N-gramas para representar os tweets, *features* comportamentais para caracterizar o consumo de álcool e *features* semânticas para fornecer contexto. Por outro lado, *Drunk2Vec* combina representações densas, na forma de *word embeddings*, e *features* semânticas para fornecer contexto aos tweets. Ambos os métodos abrangem várias etapas adicionais que lidam com: a) pré-processamento e correção de tweets com erros ortográficos devido à influência do álcool; b) engenharia adicional de *features* (símbolos) (por exemplo, erros, sentimentos); c) seleção e/ou redução de *features* para lidar com a esparcialidade e a alta dimensionalidade; d) integração de *features* (textuais, se-

mânticas ou distribucionais). É importante destacar que os métodos são independentes do algoritmo de classificação, ou seja, podem ser utilizados em conjunto com diferentes classificadores. As seções, a seguir, descrevem detalhadamente cada um dos métodos.

Figura 5.1: *Features* extraídas usando *Drunk2Symbol* e *Drunk2Vec*



Fonte: Elaborado pelo autor

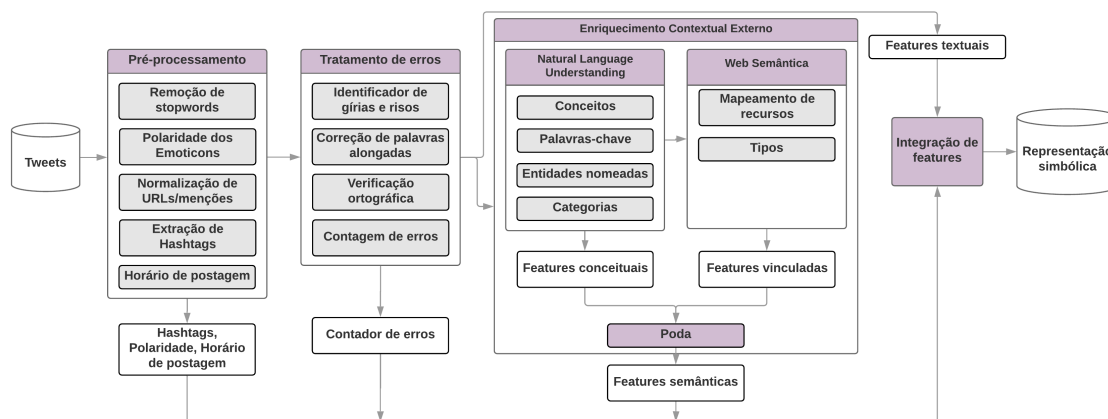
5.2 *Drunk2Symbol*

Drunk2Symbol é um método que complementa os tweets com *features* resultantes do enriquecimento contextual externo e *features* que caracterizam a embriaguez. O enriquecimento semântico fornece contexto para textos esparsos e que contêm ruídos, como ocorre em tweets, usando fontes externas de conhecimento (ferramentas *Natural Language Understanding* (NLU) e Web Semântica). *Drunk2Symbol* também extrai *features* que caracterizam o comportamento de embriaguez, como a presença de erros de digitação, sentimentos e o horário de postagem. Desta forma, *Drunk2Symbol* é um extrator de *features* simbólicas que produz bases de dados de treinamento para a classificação de mensagens de texto relacionadas ao abuso de álcool.

Drunk2Symbol é um processo composto por cinco etapas, representadas na Figura 5.2. As etapas são detalhadas nesta seção, ilustradas pelo exemplo de execução na Figura 5.3.

5.2.1 Pré-processamento e extração de *features* de embriaguez

Os objetivos deste passo são realizar o pré-processamento e extrair *features* comportamentais relacionadas a embriaguez. No pré-processamento, foram realizadas ações tradicionais de limpeza e transformação sobre tweets: remoção de *stopwords*, conversão para letras minúsculas e substituição de URLs/menções por marcadores. Palavras alongadas foram normalizadas através da remoção de caracteres que ocorrem mais de duas vezes consecutivas na mesma palavra (por exemplo, *druuuunk* para *druunk*).

Figura 5.2: Visão geral do *Drunk2Symbol*

Fonte: Elaborado pelo autor

Figura 5.3: Exemplo de execução do método *Drunk2Symbol*

Linha		
1 ^a	Tweet original	Jut did a shot of wineeeee to @anna in San Jose for being my spirit guide! #love ?
2 ^a	Emoticons	❤️ (+5) Horário de postagem Night Hashtags #love
3 ^a	Tratamento de erros	Substitui winee por wine, 1 erro (jut)
4 ^a	Tweet corrigido	Jut shot wine @mention San Jose spirit guide
5 ^a	Features conceituais	
6 ^a	Palavras-chave	wine San Jose, spirit guide Entidades nomeadas San Jose
7 ^a	Conceitos	Mediumship, Spiritualism, Spirituality
8 ^a	Categorias	/food and drink/beverages/alcoholic beverages/wine, /travel/travel guides, /health and fitness/disease/cold and flu
9 ^a	Features vinculadas	
10 ^a	Resource rdf:type	http://dbpedia.org/resource/Wine - Wikidata:Q2095, DUL:FunctionalSubstance, DBpedia:Food, DBpedia:Beverage http://dbpedia.org/resource/Spirit_guide - http://dbpedia.org/resource/San_Jose_Earthquakes - owl:Thing, Wikidata:Q476028, Wikidata:Q24229398, DUL:SocialPerson, Schema:Organization, DBpedia:SportsTeam, DBpedia:SoccerClub, DBpedia:Agent
11 ^a	Poda	
12 ^a	Cortar	Manter
13 ^a	owl:Thing, Wikidata:Q476028, Wikidata:Q24229398, DUL:SocialPerson, Schema:Organization, DBpedia:SportsTeam, DBpedia:SoccerClub, DBpedia:Agent, /health and fitness/disease/cold and flu, /travel/travel guides, Spiritualism	Wikidata:Q2095, DBpedia:Food, DBpedia:Beverage
14 ^a	Integração de features	
15 ^a	jut, shot, wine, @mention, san, jose, spirit, guide, jut_shot, shot_wine, wine_@mention, @mention_san, san_jose, jose_spirit, spirit_guide, mediumship, spirituality, #love /foodanddrink/beverages/alcoholicbeverages/cocktailsandbeer, /foodanddrink/beverages/alcoholicbeverages/wine, /artandentertainment/booksandliterature/magazines, Wikidata:Q2095, DBpedia:Food, DBpedia:Beverage, DUL:FunctionalSubstance	
16 ^a	Horário de postagem: Night	Polaridade dos Emoticon: 5 Erro(s): 1

Fonte: Elaborado pelo autor

Com relação à extração de *features* comportamentais de embriaguez, foram extraídas as *hashtags* e a polaridade dos *emoticons* usando uma lista pré-definida¹ que atribui o grau de polaridade para cada *emoticon*. Por fim, extraiu-se dos metadados do tweet o horário da postagem, categorizando-o em manhã [6:00 às 14:00), tarde [14:00 às 22:00) e noite [22:00 às 06:00). A segunda linha da Figura 5.3 ilustra as *features* de embriaguez extraídas do tweet de exemplo.

¹<https://github.com/words/emoji-emotion>

5.2.2 Tratamento de erros

Usuários que estão sob influência do álcool tendem a escrever tweets com erros de digitação e erros ortográficos. A eficácia das ferramentas de enriquecimento externo pode ser reduzida na presença de termos incorretos. Desta forma, este passo tem como objetivo executar ações corretivas sempre que possível, além de contar o número de erros. O número de erros faz parte do conjunto de *features* propostas para caracterizar o comportamento relacionado a embriaguez.

Para isso, foi adaptado um validador ortográfico baseado em dicionários que lidam com gírias da Internet e palavras alongadas para desenvolver uma ferramenta automática de verificação de erros ortográficos em tweets. Assim a ferramenta executa ações corretivas, preservando palavras que caracterizam risada (kkk, haha, hehe), redes sociais (Facebook, Instagram, Twitter), gírias (LOL, OMG, ILY), alongamentos (*loove*, *druunk*) e palavras com menos de dois caracteres. Foram considerados como alongamentos palavras cuja distância de Levenshtein é igual a um (1) e o caractere distinto é igual ao anterior. Depois de identificar o alongamento, as palavras são corrigidas. Por fim, é contabilizado o número total de erros na *feature contador de erros*. A terceira e quarta linhas do exemplo de execução (Figura 5.3) exibem as correções (*winee* é corrigido para *wine*) e *contador de erros* (por exemplo, *jut* é contabilizado como erro).

5.2.3 Enriquecimento contextual externo

Para fornecer contexto aos tweets, foram combinadas duas estratégias de enriquecimento: *Natural Language Understanding* e a Web Semântica. O componente NLU mapeia os tweets para *features conceituais*, ou seja, expande o vocabulário usado para conceitos relacionados que fornecem contexto. No entanto, apesar da expansão do vocabulário, as *features conceituais* não têm necessariamente o nível apropriado de generalização para superar a esparcialidade do vocabulário de tweets. Desta forma, a etapa de Web Semântica identifica *features vinculadas* que podem generalizar as *features textuais* e *features conceituais*. Essas estratégias de enriquecimento contextual externo são complementares, pois ambas associam textos a *features* que fornecem significado e poder de generalização, tornando-se uma maneira para complementar o conteúdo curto e limitado de tweets. O enriquecimento externo tem como efeito colateral um grande número de *features* resultantes (ROMERO; BECKER, 2019), que na prática pode prejudicar o de-

sempenho da classificação. Desta forma, *Drunk2Symbol* também abrange a seleção de *features* semânticas relevantes. Os componentes são detalhados a seguir:

1) *Natural Language Understanding*: este componente tem como principal característica mapear os tweets para *features conceituais* (palavras-chave, entidades nomeadas, conceitos e categorias), ou seja, expandir o vocabulário para conceitos relacionados que fornecem contexto. As *features conceituais* identificadas são discriminantes entre os tweets e permitem a etapa de enriquecimento via Web Semântica. Para extrair *features conceituais* dos tweets, foram empregadas duas ferramentas de NLU: Open Calais² e IBM Watson³. NLU introduz *features* tais como *San Jose* (entidade nomeada), *spirit guide* (palavra-chave), *Spiritualism* (conceito) e *‘/food and drink/beverages/alcoholic beverages/wine’* (categoria) para o exemplo de execução (Figura 5.3), conforme ilustrado na sexta, sétima e oitava linha.

2) *Web Semântica*: tem como finalidade generalizar as *features conceituais*, permitindo a identificação de padrões nos tweets. As bases de conhecimento de *Linked Open Data* (LOD), como DPBedia ou Yago, descrevem recursos por meio de propriedades como *rdf:type*, *rdf:subject*, *rdf:category*, *rdf:abstract* e *rdf:label*. Tais propriedades permitem a generalização e o relacionamento entre diferentes recursos. Desta forma, foram mapeadas as *features* textuais e as *features conceituais* em recursos da base de conhecimento DBPedia. Em seguida, extraiu-se o tipo de cada conceito (*rdf:type*). Essa etapa foi realizada usando a API sparql⁴ fornecida pela DBPedia. O conjunto de tipos resultante é denominado *features vinculadas*. Algumas *features vinculadas* extraídas podem ser irrelevantes, principalmente porque são muito genéricas ou específicas. A décima linha da Figura 5.3 ilustra um trecho das *features vinculadas* obtidas para o exemplo de execução. Enquanto que *DBpedia:Beverage* fornece contexto relevante, *thing* é muito genérico e *Wikidata:Q476028* é muito específico. Também foram realizados experimentos explorando a categoria (*rdfs:subClassOf*) dos recursos da DBPedia, mas essa alternativa apresentou resultados inferiores e, portanto, não são descritos nesta dissertação.

3) *Poda*: O enriquecimento contextual externo retorna um grande conjunto de *features*, mas apenas um número limitado delas é realmente relevante para a tarefa de classificação. Conceitos, palavras-chave e categorias extraídas usando as ferramentas NLU podem expandir o número de *features* irrelevantes (por exemplo, *Spiritualism*). Os recursos da DBPedia são descritos sob várias perspectivas, nem todas relacionadas ao

²<http://www.opencalais.com/>

³<https://www.ibm.com/watson/services/natural-language-understanding/>

⁴<https://dbpedia.org/sparql>

problema em questão. Por exemplo, os nomes dos coquetéis (*cocktail*) podem ser generalizados para bebidas alcoólicas, mas também para conceitos muito genéricos (por exemplo, *owl:Thing*) ou específicos (por exemplo, *Wikidata:Q476028*). Além disso, *features conceituais* e *features vinculadas* podem ser redundantes (por exemplo, */food and drink/beverages/alcoholic beverages/wine* vs. *DBpedia:Beverage*). O número excessivo de *features* irrelevantes pode degradar o desempenho do classificador e, portanto, é necessário aplicar um algoritmo de seleção de *features*. Portanto, foi implementado o algoritmo InfoGain para manter apenas as *features conceituais* e as *features vinculadas* discriminantes. O InfoGain calcula a entropia entre a presença da *feature* e a classe alvo, selecionando apenas as *features* relacionadas ao resultado. O algoritmo InfoGain foi executado com a configuração padrão sobre as *features conceituais* e *features vinculadas* de todos os tweets da base de dados. O conjunto final de *features* relevantes é denominado *features semânticas*. Na décima terceira linha do exemplo de execução (Figura 5.3), este componente removeu *features* muito específicas (*Wikidata:Q476028*), genéricas (*owl:Thing*), irrelevantes (*/travel/travel guides*) ou redundantes (*/food and drink/beverages/alcoholic beverages/wine* vs. *DBpedia:Beverage*).

5.2.4 Integração de *features*

O papel deste passo é transformar e combinar *features* textuais, *features semânticas* e *features* comportamentais de embriaguez (*hashtags*, presença de erros, sentimentos dos *emoticons* e o horário de postagem) em uma representação simbólica consolidada, de modo a compor uma base de dados de treinamento para classificação do consumo excessivo de álcool. As *features* textuais extraídas dos tweets corrigidos foram representadas através de bigramas (uni-gramas e bigramas), uma vez que N-gramas são mais representativos do que palavras individuais. Todas as *features semânticas* (conceitos, palavras-chave, entidades nomeadas, categorias e tipos da DBPedia) e as *features* comportamentais de embriaguez (*hashtags*, polaridade, presença de erros) são representadas usando BoW.

Por fim, todas as *features* são combinadas para compor o conjunto de dados de treinamento, conforme ilustrado na décima quinta e décima sexta linha do exemplo de execução (Figura 5.3). Essa representação simbólica pode ser processada por qualquer algoritmo de aprendizado de máquina supervisionado para classificar textos bêbados.

5.3 *Drunk2Vec*

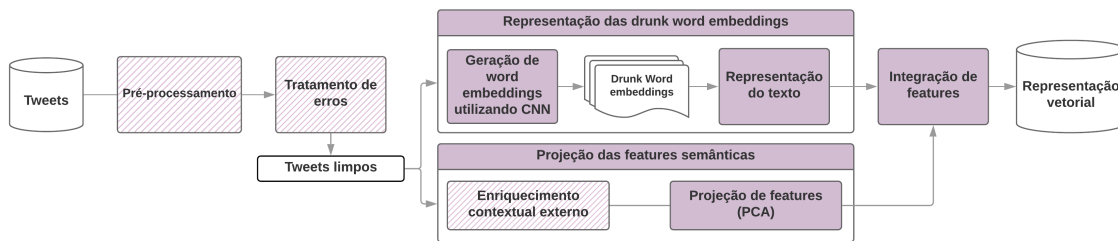
Drunk2Vec é um método que cria representações vetoriais de tweets integrando o contexto fornecido pelo vocabulário utilizado, na forma de *word embeddings* específicas para o domínio, com projeções das *features semânticas* produzidas com o método *Drunk2Symbol*. *Drunk2Vec* combina as duas estratégias de enriquecimento contextual com base na premissa de que ambas são complementares: enquanto o enriquecimento contextual externo expande e generaliza o vocabulário, lidando com a esparcialidade, a semântica distribucional lida com as idiosincrasias do vocabulário de textos bêbados, uma vez que aprende representações que refletem o vocabulário utilizado.

O núcleo do *Drunk2Vec* é o uso de uma rede neural convolucional para aprender de maneira supervisionada uma infinidade de relacionamentos entre termos que podem ser derivados do contexto de mensagens de textos bêbados e representá-los como *word embeddings* específicas do domínio. Essa abordagem supera as limitações relacionadas ao uso de algoritmos não supervisionados genéricos (por exemplo, *Word2Vec* ou *Glove*), que normalmente exigem grandes volumes de instâncias de treinamento para produzir bons resultados, bem como *word embeddings* pré-treinadas, nas quais é possível enfrentar problemas de OOV. A arquitetura da CNN resolve esses problemas porque as convoluções buscam padrões locais no conjunto de dados, representando-os através de vetores.

A interpolação direta de *features semânticas* e *word embeddings* resultaria em uma matriz esparsa e de alta dimensionalidade, o que pode prejudicar o desempenho da classificação. Assim, *Drunk2Vec* gera representações densas das *features semânticas* resultantes do *Drunk2Symbol* usando a Análise de Componentes Principais (PCA). O PCA cria uma projeção equivalente que representa as *features semânticas* em um formato compacto, mantendo o conhecimento obtido por meio do enriquecimento contextual externo, sem introduzir esparcialidade no conjunto de dados.

O método *Drunk2Vec* abrange cinco etapas, conforme ilustrado na Figura 5.4. As etapas de Pré-processamento, Tratamento de Erros e Enriquecimento Contextual Externo são semelhantes às aplicadas no *Drunk2Symbol*. Detalhes adicionais são fornecidos no restante desta seção e ilustrados usando um exemplo de execução (Figura 5.5).

Figura 5.4: Visão geral do *Drunk2Vec*



Fonte: Elaborado pelo autor

Figura 5.5: Exemplo de execução do método *Drunk2Vec*

Linha																																																			
1ª	Tweet original	Jut did a shot of wineeeee to @anna in San Jose for being my spirit guide! #love ?																																																	
2ª	Pré-processamento e tratamento de erros	jut did a shot of wine to @mention in san jose for being my spirit guide! #love ?																																																	
3ª	Geração de drunk word embeddings																																																		
4ª	Drunk word embeddings shot is similar to: shots, killed, hit, shooting, and after wine is similar to: beer, bottle, drink, tasting, and coffee mention is similar to: mention, mentions, tweet, retweet, follow, and add spirit is similar to: god, our, heart, great, and life	Projeção das features semânticas Features semânticas Wikidata:Q2095, DBpedia:Food, DBpedia:Beverage																																																	
5ª	Representação do texto	Projeção com PCA 0.981 0.123																																																	
	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Palavra</th> <th>V1</th> <th>V2</th> <th>V3</th> <th>V4</th> <th>...</th> <th>V100</th> </tr> </thead> <tbody> <tr> <td>jut</td> <td>Uu</td> <td>2.669</td> <td>-2.954</td> <td>1.883</td> <td></td> <td>2.394</td> </tr> <tr> <td>did</td> <td>0.849</td> <td>-0.603</td> <td>-2.953</td> <td>2.184</td> <td></td> <td>-0.399</td> </tr> <tr> <td>shot</td> <td>-1.137</td> <td>0.088</td> <td>2.200</td> <td>-2.937</td> <td></td> <td>-0.927</td> </tr> <tr> <td>...</td> <td>...</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>guide</td> <td>-1.049</td> <td>1.224</td> <td>1.615</td> <td>-0.034</td> <td></td> <td>0.482</td> </tr> <tr> <td>Média do tweet</td> <td>-0.538</td> <td>-1.576</td> <td>0.361</td> <td>1.632</td> <td></td> <td>12.45</td> </tr> </tbody> </table>	Palavra	V1	V2	V3	V4	...	V100	jut	Uu	2.669	-2.954	1.883		2.394	did	0.849	-0.603	-2.953	2.184		-0.399	shot	-1.137	0.088	2.200	-2.937		-0.927						guide	-1.049	1.224	1.615	-0.034		0.482	Média do tweet	-0.538	-1.576	0.361	1.632		12.45	
Palavra	V1	V2	V3	V4	...	V100																																													
jut	Uu	2.669	-2.954	1.883		2.394																																													
did	0.849	-0.603	-2.953	2.184		-0.399																																													
shot	-1.137	0.088	2.200	-2.937		-0.927																																													
...	...																																																		
guide	-1.049	1.224	1.615	-0.034		0.482																																													
Média do tweet	-0.538	-1.576	0.361	1.632		12.45																																													
6ª	Integração de features	-0.538 -1.576 0.361 ... 1.632 12.4 0.981 0.123																																																	

Fonte: Elaborado pelo autor

5.3.1 Pré-processamento e tratamento de erros

Essas etapas são semelhantes às descritas no método *Drunk2Symbol* (Seções 5.2.1 e 5.2.2). As principais diferenças são: a) *stopwords* e *hashtags* não são removidas; b) features comportamentais de embriaguez não são extraídas; c) tweets com 5 ou menos caracteres são descartados. O pré-processamento e o tratamento de erros são importantes para corrigir e limpar os tweets, que são usados como entrada de texto para as próximas etapas. Além de melhorar o desempenho da NLU (Seção 5.2.3), essas etapas também contribuem para a qualidade das *word embeddings* resultantes, evitando que termos idênticos, porém com erros de ortografia ou variações alongadas sejam representados por vetores distintos. O resultado dessas ações é ilustrado na segunda linha do exemplo de execução (Figura 5.5).

5.3.2 Representação das *drunk word embeddings*

Os objetivos desta etapa são gerar *drunk word embeddings* e usá-las para representar o contexto dos tweets. Esses componentes são detalhados a seguir.

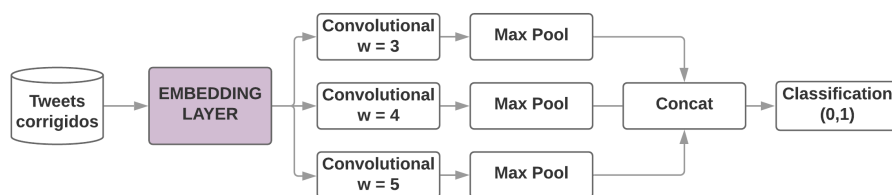
1) *Geração de drunk embeddings*: *word embeddings* capturam contexto aprendendo representações que refletem o uso das palavras, de modo que termos diferentes empregados em um contexto semelhante tendem a ser similares. Alguns exemplos são exibidos na quarta linha do exemplo de execução (Figura 5.5). O vetor que representa o termo *wine* é semelhante as *word embeddings* dos termos que representam bebidas (*beer, drink, coffee, bottle*), enquanto o vetor da palavra *mention* é semelhante aos termos usados para denotar interações nas redes sociais (por exemplo, *mentions, tweet, retweet, follow*). *Drunk2Vec* aprende *word embeddings* a partir de um corpus de domínio específico usando uma CNN, já que não se dispõe de um grande conjunto de dados para aprender *word embeddings* usando algoritmos genéricos. Também não foram empregadas *word embeddings* pré-treinadas sob a suposição de que os aspectos específicos da linguagem utilizada por pessoas sob influência do álcool não são representados por eles (BADJATIYA et al., 2017).

Para gerar *word embeddings*, foi treinada uma rede neural CNN usando um conjunto de dados rotulado, de modo que os vetores de *word embeddings* são aprendidos em conjunto com a tarefa principal, ou seja, da mesma maneira que a rede neural aprende seus pesos. Esse processo de aprendizado considera a classe alvo para ajustar os vetores de *embeddings*. Optou-se por aplicar CNNs com vários filtros de tamanho (3,4,5) pela sua capacidade de combinar diferentes padrões locais identificados pelas camadas convolucionais, fornecendo generalizações e otimizando seu uso em textos curtos.

A CNN é uma adaptação da proposta em (KIM, 2014) e é composta por cinco camadas, ilustradas na Figura 5.6. A camada de *embedding* converte palavras em *word embeddings*. Nesta abordagem, todos os vetores de *embeddings* são inicializados aleatoriamente e modificados durante o treinamento. Na segunda camada, são usadas três camadas convolucionais com filtros de diferentes tamanhos para extrair sequências e identificar padrões no conjunto de dados, onde o tamanho define o número de palavras próximas usadas para extrair os padrões locais. A terceira camada identifica os recursos mais importantes da sequência através de operações de *pooling*. A quarta camada mescla os resultados das operações de *pooling* em um vetor único de *features*. A última camada aplica operações de regularização, densas e de normalização, evitando *overfitting* e executando a tarefa de classificação. Essa rede neural foi executada e, ao final do processo, foram extraídos os pesos da camada de *embeddings* da rede neural, ou seja, os vetores das palavras com os pesos ajustados pela rede neural. Essa camada representa cada palavra por um vetor de 100 dimensões. O código fonte da rede neural CNN é disponibilizado em um repositório

público⁵.

Figura 5.6: Arquitetura da CNN para aprender *drunk word embeddings*



Fonte: Elaborado pelo autor

2) *Representação do texto*: dado um conjunto de *drunk word embeddings*, a representação de cada tweet foi composta utilizando a média de todas as *word embeddings* que representam as palavras que o tweet contém. Essa escolha foi tomada pelos seguintes motivos: (a) uma representação unidimensional dos tweets permite torná-los independente do algoritmo de classificação; e (b) permite a integração de *word embeddings* e *features semânticas*. Também foram realizados experimentos preliminares com sequências de palavras, mas essa alternativa apresentou resultados inferiores e, portanto, não são descritos nesta dissertação. A quinta linha no lado direito da Figura 5.5 ilustra o vetor médio para o exemplo de execução.

5.3.3 Projeção das *features semânticas*

Esta etapa visa extrair as *features semânticas* e representá-las usando uma representação densa com baixa dimensionalidade. Para extrair as *features semânticas*, foi utilizada a etapa de Enriquecimento Contextual Externo do método *Drunk2Symbol*, conforme descrito na Seção 5.2.3.

As *features semânticas* devem ser integradas às *word embeddings* criadas para representar o conteúdo textual. Uma alternativa é representar essas *features* por meio da codificação *one-hot*, mas o resultado seria uma matriz esparsa, em que o tamanho de cada vetor é igual ao tamanho do vocabulário de todas as *features semânticas*. Para resolver esse problema, propõe-se o uso do PCA para codificar as *features semânticas* usando representações densas. O PCA cria uma projeção equivalente que representa as *features semânticas* em um formato compacto, mantendo o conhecimento obtido através do enriquecimento contextual externo. O resultado do PCA sobre as *features semânticas* é ilustrado no lado direito das linhas 4 e 5 do exemplo de execução (Figura 5.5), usando

⁵<https://gist.github.com/MarcosGrzeca/435cbcbd451c4a31831520028bca8a6e>

duas dimensões.

5.3.4 Integração de *features*

Esta etapa concatena o vetor médio que representa os tweets com as principais componentes das respectivas *features semânticas*, criando uma representação de baixa dimensionalidade que se beneficia do poder das duas técnicas de enriquecimento contextual. A última linha do exemplo de execução (Figura 5.5) ilustra esta etapa. O resultado é um conjunto de dados composto por representações vetoriais que pode ser usado para treinar diferentes algoritmos de classificação, incluindo redes neurais profundas.

6 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta os experimentos realizados para avaliar os métodos *Drunk2Symbol* e *Drunk2Vec*, propostos no Capítulo 5. Inicialmente são descritos os objetivos dos experimentos, as bases de dados e as ferramentas utilizadas. Por fim, é descrita a metodologia e os resultados dos experimentos.

6.1 Objetivos

Quatro experimentos foram desenvolvidos para avaliar o desempenho dos métodos propostos para identificar tweets bêbados. Os objetivos específicos de cada experimento são detalhados a seguir:

- *Experimento #1*: verificar o desempenho do método *Drunk2Symbol* e avaliar as contribuições das *features* propostas (*features semânticas* e *features comportamentais de embriaguez*);
- *Experimento #2*: avaliar o desempenho do método *Drunk2Vec*, assim como a eficácia da integração de ambas as estratégias de enriquecimento contextual;
- *Experimento #3*: avaliar se a adoção de um conjunto de classificadores é uma solução eficaz para integrar as duas estratégias de enriquecimento contextual;
- *Experimento #4*: analisar a eficácia da rede neural CNN para gerar *word embeddings* específicas do domínio, comparando-a com a geração de vetores usando *Word2Vec* e *word embeddings* pré-treinadas.

6.2 Bases de dados

Os experimentos foram executados utilizando todas as bases de dados descritas no Capítulo 4. Cada base de dados representa um comportamento diferente do consumo de álcool e a utilização do Twitter:

- *DS1-drunk-mention*: é composta de tweets que mencionam o consumo de álcool. No entanto, esta base de dados também pode conter tweets relacionados a notícias ou anúncios vinculados ao consumo de álcool;
- *DS1-drunk-drinking*: representa os tweets nos quais o usuário está consumindo

álcool;

- *DS1-drunk-drink-now*: contém tweets em que o usuário está tuitando enquanto está alcoolizado;
- *DS2-keywords*: engloba uma amostra aleatória de tweets classificados por palavras-chave;
- *DS3-drinking-ext*: base de dados estendida de *DS1-drunk-drinking*, ou seja, contém tweets nos quais o usuário está consumindo álcool.

Tabela 6.1: Tamanho das bases de dados

Base de dados	#Tweets bêbados	#Tweets
DS1-drunk-mention	2.236	3.994
DS1-drunk-drinking	1.374	1.777
DS1-drunk-drink-now	651	1.054
DS2-keywords	5.000	11.792
DS3-drinking-ext	3.147	4.456

A Tabela 6.1 detalha o tamanho das bases de dados e o número de tweets bêbados. As bases de dados *DS1-drunk-drinking*, *DS1-drunk-drink-now* e *DS3-drinking-ext* são mais relevantes para a classificação de tweets bêbados, uma vez que a primeira base de dados é anotada pela mera menção ao álcool (incluindo anúncios) e *DS2-keywords* é uma amostra aleatória, assim *DS1-drunk-mention* e *DS2-keywords* não avaliam se o usuário está alcoolizado.

6.3 Ferramentas e métricas de avaliação

A linguagem de programação *R* foi utilizada para desenvolver os métodos de classificação. Também foram utilizadas várias bibliotecas complementares, das quais destacam-se a *Caret*¹ e a *Keras*². *Caret* disponibiliza algoritmos de aprendizado de máquina, enquanto a *Keras* possibilita a utilização do aprendizado profundo e *word embeddings*. O enriquecimento contextual externo foi implementado através da linguagem de programação *PHP*, que se integrou com as APIs do *Calais*, *Watson* e *DBPedia*.

Foram experimentados três algoritmos de classificação: *Support Vector Machine* (SVM) Poly, *eXtreme Gradient Boosting* (XGBoost) e *Deep Neural Network* (DNN). Esses são os algoritmos mais usados em soluções vencedoras nas competições do *Kaggle*³.

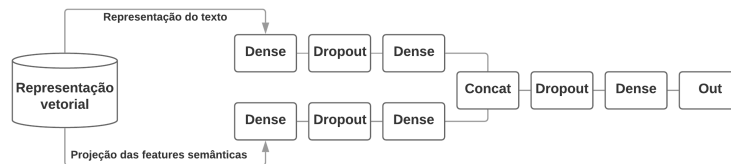
¹<http://caret.r-forge.r-project.org/>

²<https://keras.rstudio.com/>

³<https://www.kaggle.com/bigfatdata/what-algorithms-are-most-successful-on-kaggle>

A arquitetura da DNN é composta por camadas densas, *dropout* e a função objetivo *binary_crossentropy*, conforme ilustrado na Figura 6.1. Para cada método, os algoritmos foram treinados por dez (10) vezes, registrando a média dos resultados para a classe positiva (*bêbado*). A classificação foi executada reservando 80% dos dados para treinamento e 20% para testes, com validação cruzada de 5 partições.

Figura 6.1: Arquitetura da DNN



Fonte: Elaborado pelo autor

Por fim, a análise dos resultados baseou-se nas métricas de avaliação definidas na Seção 2.5, ou seja, precisão, revocação e medida F_1 . Para verificar se os resultados são estatisticamente significativos, foi realizado o t-test pareado utilizando o nível de significância padrão $\alpha = 0,05$.

6.4 Experimento #1: *Drunk2Symbol*

O objetivo deste experimento é avaliar o desempenho do método *Drunk2Symbol* com aprendizado de máquina para classificar tweets. A suposição analisada é que as *features* de enriquecimento contextual externo e as *features* comportamentais de embriaguez ajudam a melhorar a identificação de mensagens de texto sob influência do álcool com melhorias estatisticamente significativas.

6.4.1 Metodologia

A primeira etapa deste experimento envolveu a execução do método *Drunk2Symbol* para cada base de dados. Em seguida, foram testados os algoritmos de classificação *SVM Poly* (*Drunk2Symbol-SVM*) e *XGBoost* (*Drunk2Symbol-XGBoost*). Para confirmar a eficácia do método, comparou-se os resultados obtidos com o método proposto por (HOSSAIN et al., 2016), utilizada como *baseline* por alcançar a melhor medida F_1 entre os trabalhos relacionados. Para obter as métricas de desempenho do classificador *baseline*, o trabalho descrito em

Hossain et al. (2016) foi reproduzido. Por fim, foram analisadas as trinta (30) variáveis mais influentes para cada classificador, com o intuito de avaliar a contribuição do enriquecimento contextual externo e das *features* comportamentais.

A Tabela 6.2 sumariza as *features* utilizadas neste experimento. Inicialmente, são descritas as quantidades de *features* textuais, *hashtags* e comportamentais. Em seguida, são detalhadas as quantidades de *features* conceituais e vinculadas extraídas. Por fim, são contabilizados os totais de *features* conceituais e vinculadas após a aplicação do componente de poda. É importante destacar que para as bases de dados *DS1-drunk-mention*, *DS1-drunk-drinking* e *DS1-drunk-drink-now* foi utilizado o mesmo conjunto de *features* vinculadas e conceituais. Esse conjunto corresponde a união de todas as *features* vinculadas e conceituais relevantes para *DS1-drunk-mention*, *DS1-drunk-drinking* e *DS1-drunk-drink-now*.

Tabela 6.2: *Features* utilizadas no Experimento #1: *Drunk2Symbol*

Base de dados	Textuais	Hashtags	Comportamentais	Conceituais		Vinculadas	
				Sem poda	Com poda	Sem poda	Com poda
DS1-drunk-mention	30452	881	2	9746	51	7181	48
DS1-drunk-drinking	14969	450	2	4928	51	4124	48
DS1-drunk-drink-now	9360	305	2	3268	51	2868	48
DS2-keywords	7685	390	2	9552	84	5251	186
DS3-drinking-ext	32363	2431	2	15831	14	8806	44

6.4.2 Resultados

A seguir são detalhados os resultados obtidos pelo método *Drunk2Symbol* na classificação de tweets bêbados e avaliada a contribuição das *features* propostas pelo método.

6.4.2.1 Desempenho na classificação de tweets bêbados

A Tabela 6.3 apresenta os resultados para a medida F_1 , precisão e revocação para o *baseline*, assim como para *Drunk2Symbol-SVM* e *Drunk2Symbol-XGBoost*, ou seja, os algoritmos SVM e XGBoost treinados com as *features* resultantes do método *Drunk2Symbol* para cada base de dados. Os resultados destacados em negrito com o símbolo (v) indicam que as diferenças entre *Drunk2Symbol* (SVM ou XGBoost) e o *baseline* são estatisticamente significativas. Nos demais testes, não há diferença estatística entre os resultados.

Drunk2Symbol-SVM supera o *baseline* na medida F_1 em todos os experimentos, com ganhos variando entre 1,4 e 9,2 pontos percentuais (pp). Todas as melhorias na

Tabela 6.3: Resultados do Experimento #1

Base de dados	Hossain			Drunk2Symbol-SVM			Drunk2Symbol-XGBoost		
	F ₁	P	R	F ₁	P	R	F ₁	P	R
DS1-drunk-mention	88,360	89,047	87,674	89,834	92,151 v	87,517	89,461	91,574 v	87,484
DS1-drunk-drinking	85,504	81,662	89,343	89,057 v	81,398	96,715 v	86,381	81,734	91,634
DS1-drunk-drink-now	78,788	76,191	81,392	88,037 v	80,892 v	95,182 v	80,916	76,885	85,486 v
DS2-keywords	78,818	81,705	76,130	80,767 v	86,714 v	75,600	79,746	84,542 v	75,477
DS3-drinking-ext	79,944	75,343	85,147	82,854 v	75,154	92,349 v	81,355 v	76,527 v	86,861 v
Média	82,283	80,790	83,937	86,110	83,262	89,473	83,572	82,252	85,389

medida F_1 são estatisticamente significativas, exceto para *DS1-drunk-mention*. Os resultados mais significativos foram obtidos na revocação, onde as melhorias foram de 7,3 pp, 13,7 pp e 7,2 pp em situações reais de mensagens de textos bêbados (*DS1-drunk-drinking*, *DS1-drunk-drink-now* e *DS3-drinking-ext*, respectivamente). Para as bases de dados que mencionam álcool (*DS1-drunk-mention* e *DS2-keywords*), a revocação foi equivalente ao *baseline*. Também foram alcançadas melhorias na precisão para as bases de dados *DS1-drunk-mention*, *DS1-drunk-drink-now* e *DS2-keywords*, variando entre 2,4 pp e 5 pp.

Drunk2Symbol-XGBoost também obteve melhorias em comparação ao *baseline*. A medida F_1 é superior para todas as bases de dados, mas a única melhoria estatisticamente significativa foi observada para *DS3-drinking-ext*. No que se refere a revocação, *Drunk2Symbol-XGBoost* obteve resultados superiores para todas as bases de dados que representam situações reais de mensagens de textos bêbados (*DS1-drunk-drinking*, *DS1-drunk-drink-now* e *DS3-drinking-ext*), sendo que em dois casos as melhorias são estatisticamente significativas. *Drunk2Symbol-XGBoost* também obteve melhorias estatisticamente significativas na precisão nas bases de dados que mencionam álcool (*DS1-drunk-mention* e *DS2-keywords*) e no conjunto de dados *DS3-drinking-ext*, que correspondem as três maiores bases de dados deste trabalho.

Em resumo, os algoritmos de classificação treinados usando as *features* do *Drunk2Symbol* superam o *baseline* na identificação de textos bêbados, principalmente na revocação devido à expansão e generalização do vocabulário. Entre os dois algoritmos, o *Drunk2Symbol-SVM* apresentou a maior média para todas as métricas de avaliação, especialmente para situações de textos bêbados. Considerando todas as métricas e bases de dados, o *Drunk2Symbol-SVM* superou o *Drunk2Symbol-XGBoost* em 86% dos testes. Como o *Drunk2Symbol-SVM* é o melhor resultado do Experimento #1, adotou-se como *baseline* no Experimento #2 (Seção 6.5).

6.4.2.2 Contribuição das *features* propostas

Para avaliar a contribuição do enriquecimento contextual externo e das *features* comportamentais nos resultados, foram analisadas as trinta (30) variáveis⁴ mais importantes para o classificador `Drunk2Symbol-SVM`. A avaliação foi realizada para todas as bases de dados e os resultados são exibidos na Tabela 6.4. As *features* semânticas representam entre 60-73% das *features* mais relevantes para os classificadores. Essa alta proporção entre as 30 variáveis mais importantes demonstra a importância do enriquecimento contextual externo para o desempenho da tarefa de classificação e o papel complementar de ambas as estratégias de enriquecimento. Ao analisar as *features* semânticas para *DS1-drunk-drinking*, verifica-se que as melhorias foram possíveis, pois diferentes palavras no texto que se referem a bebidas alcoólicas (por exemplo, */food and drink/beverages alcoholic/wine cocktails, yago.Drug103247620, yago.WikicatDrugs*) foram agrupadas de acordo com seu significado contextual. Além disso, a presença de erros foi a segunda *feature* mais importante para *DS1-drunk-drink-now* e a décima nona mais influente para *DS2-keywords*. O horário de postagem também aparece entre as mais relevantes para *DS2-keywords*.

Tabela 6.4: Trinta *features* mais importantes para `Drunk2Symbol-SVM`

Base de dados	Vinculadas	Conceituais	% no top-30	Presença de erros	Horário de postagem
DS1-drunk-mention	17	4	70%		
DS1-drunk-drinking	18	4	73%		
DS1-drunk-drink-now	15	4	63%	X	
DS2-keywords	13	5	60%	X	X
DS3-drinking-ext	14	7	70%		

Conclui-se que as melhorias significativas produzidas pelo método *Drunk2Symbol-SVM* são relacionadas às *features* semânticas. As *features* semânticas auxiliam na identificação de padrões nos tweets e melhoram a detecção de textos bêbados. As *features* vinculadas destacam-se entre as *features* semânticas, pois apresentam a maior representatividade entre as *features* mais importantes para a tarefa de classificação. As *features* comportamentais de embriaguez foram relevantes em apenas duas bases de dados e, portanto, seu efeito é limitado.

⁴<https://www.rdocumentation.org/packages/caret/versions/6.0-79/topics/plot.varImp.train>

6.5 Experimento #2: *Drunk2Vec*

Este segundo experimento examina os benefícios do uso do método *Drunk2Vec* e possui dois objetivos específicos: *i*) avaliar a contribuição do *Drunk2Vec* para classificar mensagens de textos bêbados; *ii*) analisar as melhorias obtidas combinando enriquecimento contextual externo e *word embeddings*. A primeira suposição avaliada é que o *Drunk2Vec* pode melhorar a detecção de palavras semelhantes, uma vez que *word embeddings* se concentram no significado das palavras e não na ortografia. Também se analisa a suposição de que a integração entre *features semânticas* e *word embeddings* produz melhorias em comparação com cada estratégia de enriquecimento isolada.

6.5.1 Metodologia

O método *Drunk2Vec* foi executado para cada base de dados em conjunto com três diferentes algoritmos de classificação: SVM Poly (*Drunk2Vec-SVM*), XGBoost (*Drunk2Vec-XGBoost*) e Deep Neural Network (*Drunk2Vec-DNN*). É importante destacar que o classificador *Drunk2Symbol-SVM* é utilizado como *baseline* para este experimento, devido aos melhores resultados obtidos no Experimento #1. Para avaliar a contribuição de cada estratégia de enriquecimento contextual, o *Drunk2Vec* foi alterado para não incluir as *features semânticas*, de modo a explorar apenas o contexto fornecido pelas *word embeddings*. Em seguida, foram aplicados os mesmos algoritmos de classificação mencionados anteriormente, criando os classificadores *Drunk2Vec-DNN*, *Drunk2Vec-SVM* e *Drunk2Vec-XGBoost*. Por fim, foi apurada a diferença entre o desempenho médio dos classificadores correspondentes para todas as métricas de avaliação.

Para gerar *word embeddings* específicas do domínio, foi executada a rede neural CNN para cada base de dados por dez (10) épocas. Também foi parametrizado o tamanho do lote igual a 64 e reservado 80% dos dados para treinamento. No final da execução, foi extraída a camada que representa as *word embeddings*. Experimentos preliminares indicaram que um vetor de palavras com 100 dimensões é suficiente para representar as *word embeddings*, uma vez que a utilização de 150 e 200 dimensões apresentou resultados similares.

Em relação a projeção das *features semânticas*, foram desenvolvidos experimentos que demonstraram que 15% das principais componentes são suficientes para representar

o conjunto de *features semânticas*. Também foi testada a incorporação de *features semânticas* usando a estratégia de codificação *one-hot*, mas os resultados foram inferiores e não são reportados neste trabalho.

A metodologia de testes é a mesma utilizada no Experimento #1 para todos os classificadores, ou seja, dez execuções reservando 80% dos dados para treinamento e 20% para testes. Em relação ao classificador DNN, o mesmo foi treinado por cinco (5) épocas, das quais escolheu-se a época ótima para avaliar o modelo. Uma época é considerada ótima quando a métrica *validation loss* é minimizada.

6.5.2 Resultados

A seguir são detalhados os resultados obtidos pelo método *Drunk2Vec* na classificação de tweets bêbados e avaliado o desempenho da integração entre *word embeddings* e *features semânticas*.

6.5.2.1 Desempenho na classificação de tweets bêbados

A Tabela 6.5 apresenta a média dos resultados para a medida F_1 , precisão e revocação do *baseline* (*Drunk2Symbol-SVM*), assim como para *Drunk2Vec-DNN*, *Drunk2Vec-XGBoost* e *Drunk2Vec-SVM*, ou seja, os algoritmos DNN, SVM e XGBoost treinados usando as *features* extraídas pelo *Drunk2Vec* para cada base de dados. Na Tabela 6.5, o símbolo (*) representa que o *baseline* é estatisticamente superior, enquanto que o símbolo (v), em negrito, indica que as melhorias obtidas pelo classificador proposto (*Drunk2Vec-DNN*, *Drunk2Vec-SVM* ou *Drunk2Vec-XGBoost*) são estatisticamente significativas.

As melhorias mais significativas foram obtidas na precisão, com melhorias que variam entre 1,7 pp e 11,4 pp. Todas as melhorias na precisão são estatisticamente significativas, exceto para *DS2-keywords* utilizando o classificador *Drunk2Vec-DNN*. Em relação a revocação, o *Drunk2Vec* apresenta melhorias de até 11 pp para as bases de dados que mencionam álcool (*DS1-drunk-mention* e *DS2-keywords*). No entanto, para a revocação, os resultados foram estatisticamente inferiores ou semelhantes para todas as bases de dados que representam situações reais de mensagens de textos bêbados. Finalmente, o *Drunk2Vec* também melhora a medida F_1 entre 0,8 pp e 7,8 pp em todas as bases de dados, principalmente devido às melhorias na precisão. As melhorias obtidas

Tabela 6.5: Resultados do Experimento #2

Base de dados	Drunk2Symbol-SVM			Drunk2Vec-DNN			Drunk2Vec-SVM			Drunk2Vec-XGBoost		
	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R
DS1-drunk-mention	89,834	92,151	87,517	93,919 v	93,979 v	93,880 v	94,453 v	94,661 v	94,257 v	93,318 v	93,925 v	92,734 v
DS1-drunk-drinking	89,057	81,398	96,715	93,072 v	92,815 v	93,362 *	93,435 v	91,036 v	95,985	92,497 v	90,628 v	94,490 *
DS1-drunk-drink-now	88,037	80,892	95,182	89,107	90,663 v	87,733 *	90,080 v	88,704 v	91,620 *	88,832	89,344 v	88,437 *
DS2-keywords	80,767	86,714	75,600	85,729 v	86,372	85,264 v	88,695 v	89,869 v	87,563 v	88,226 v	88,904 v	87,565 v
DS3-drinking-ext	82,854	75,154	92,349	88,437 v	86,102 v	90,923	88,440 v	85,474 v	91,635	88,016 v	86,348 v	89,759 *
Média	86,110	83,262	89,473	90,053	89,986	90,233	91,021	89,949	92,212	90,178	89,830	90,597

na medida F_1 são estatisticamente significativas, exceto para *DS1-drunk-drink-now* com *Drunk2Vec-DNN* e *Drunk2Vec-XGBoost*.

Conclui-se que os algoritmos de classificação treinados de acordo com as *features* do *Drunk2Vec* superam os bons resultados alcançados pelo *Drunk2Symbol*. Os melhores resultados são obtidos na precisão capturando as especificidades da linguagem empregada em textos bêbados. Em alguns casos, também ocorrem melhorias na revocação. *Drunk2Vec-DNN* produz índices de precisão superiores aos demais classificadores para situações de tweets bêbados, embora à custa da revocação. Por outro lado, *Drunk2Vec-SVM* obteve a mais alta medida F_1 e possui revocação semelhante ao experimento anterior, confirmando o poder do algoritmo SVM Poly em identificar corretamente a classe positiva (bêbado). *Drunk2Vec-XGBoost* apresenta melhorias na precisão às custas da revocação para as situações de mensagens de textos bêbados. Esses resultados confirmaram a importância das *word embeddings* para identificar termos semelhantes e melhorar a classificação de textos curtos.

6.5.2.2 Contribuição da integração entre drunk word embeddings e features semânticas

Para avaliar a contribuição do enriquecimento contextual externo, foram removidas as *features semânticas* do método *Drunk2Vec*. A Tabela 6.6 apresenta a diferença média entre os classificadores correspondentes treinados com e sem as *features semânticas*. O símbolo (v) em negrito denota que o classificador com enriquecimento contextual externo é estatisticamente superior quando comparado ao mesmo classificador sem as *features semânticas*.

Drunk2Vec produz resultados ligeiramente melhores para a medida F_1 em todas as bases de dados e classificadores, em comparação com o mesmo classificador sem essas *features*, mas a melhoria é mínima. Melhorias estatisticamente significativas apenas foram observadas para *DS2-keywords*. Conclui-se que a arquitetura *Drunk2Vec* não é eficaz na integração de diferentes abordagens de enriquecimento contextual para melhorar a classificação de tweets bêbados em situações reais de consumo.

Tabela 6.6: Melhorias obtidas usando enriquecimento contextual externo

Base de dados	Drunk2Vec-DNN			Drunk2Vec-SVM			Drunk2Vec-XGBoost		
	F ₁	P	R	F ₁	P	R	F ₁	P	R
DS1-drunken-mention	0,211	0,609	-0,184	0,334	0,353	0,312	0,292	0,294	0,294
DS1-drunken-drinking	0,004	0,012	-0,022	0,079	0,118	0,031	0,439	0,024	0,888
DS1-drunken-drink-now	0,026	0,427	-0,375	0,314	0,128	0,476	0,650	1,303	-0,341
DS2-keywords	6,561 v	7,135 v	6,095 v	5,385 v	5,465 v	5,307 v	6,477 v	6,238 v	6,702 v
DS3-drinking-ext	0,169	0,883	-0,671	0,014	0,031	0,008	0,500	0,829	0,137
Média	<i>1,394</i>	<i>1,813</i>	<i>0,969</i>	<i>1,225</i>	<i>1,219</i>	<i>1,227</i>	<i>1,672</i>	<i>1,738</i>	<i>1,536</i>

6.6 Experimento #3: *Drunk2Ensemble*

Os experimentos anteriores demonstraram que o enriquecimento contextual externo melhora o reconhecimento de textos bêbados devido à adição de *features* que incorporam significado e generalizam *features* textuais. Por outro lado, a semântica distribucional melhora a classificação de textos bêbados lidando com as idiossincrasias do vocabulário. Isso reflete em melhorias na revocação e na precisão de acordo com cada estratégia de enriquecimento contextual, respectivamente. Além disso, *Drunk2Vec* demonstrou-se ineficaz para integrar ambas as estratégias de enriquecimento contextual.

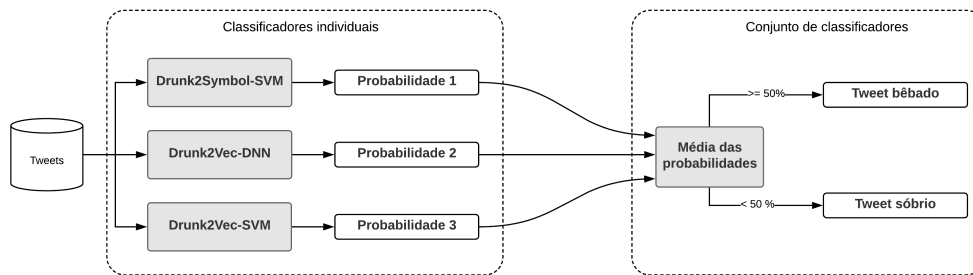
Este experimento avalia a premissa de que um conjunto de classificadores explorando ambos os métodos de enriquecimento contextual é eficaz para classificar textos bêbados e para combinar ambas as estratégias de enriquecimento contextual. Para realizar essa análise foi construído um conjunto de classificadores usando a técnica de empilhamento. A técnica combina os três melhores classificadores dos experimentos anteriores. Na última subseção deste experimento, é realizado um estudo dos casos de falhas dos classificadores propostos e descritas possíveis melhorias.

6.6.1 Metodologia

Drunk2Ensemble é composto por um conjunto de classificadores que utiliza a técnica de empilhamento para combinar as previsões dos três melhores classificadores (*Drunk2Symbol-SVM*, *Drunk2Vec-DNN* e *Drunk2Vec-SVM*), conforme ilustrado na Figura 6.2. Cada base de dados foi dividida em conjunto de treinamento (80%) e conjunto de testes (20%). Cada classificador foi treinado usando o mesmo conjunto de dados e exporta uma probabilidade entre 0 e 1 para a classe positiva. Em seguida, foi armazenada a probabilidade da classe positiva para cada classificador e tweet do conjunto de testes. Para determinar a classificação final foi calculada a média das probabilidades e

um tweet foi classificado como positivo caso a probabilidade média for igual ou superior a 50%. Esse processo foi executado por cinco (5) vezes, registrando a média da medida F_1 , precisão e revocação. Por fim, comparou-se o `Drunk2Ensemble` com o desempenho dos classificadores individuais.

Figura 6.2: Fluxo de execução do `Drunk2Ensemble`



6.6.2 Resultados

A Tabela 6.7 apresenta a média dos resultados para a medida F_1 , precisão e revocação para `Drunk2Ensemble`, bem como a diferença entre o desempenho médio do conjunto de classificadores e de cada classificador individual. As Tabelas 6.8, 6.9 e 6.10 apresentam os resultados dos testes estatísticos, de modo que os valores representados em negrito com o símbolo (v) indicam que o `Drunk2Ensemble` apresenta melhorias estatisticamente superiores. Por outro lado, os valores indicados com o símbolo (*) denotam que o classificador individual apresenta desempenho estatisticamente superior.

`Drunk2Ensemble` supera todos os outros classificadores na medida F_1 para todas as bases de dados, com melhorias significativas de até 9,3 pp. `Drunk2Ensemble` superou `Drunk2Symbol-SVM` na maioria das métricas, onde as diferenças positivas são estatisticamente superiores. Assim, a estratégia do conjunto de classificadores superou as limitações de precisão da estratégia de enriquecimento externo. Com relação a `Drunk2Vec-DNN`, é possível observar que o conjunto de classificadores melhorou a revocação sem afetar a precisão e, portanto, o conjunto de classificadores também foi uma solução eficaz nesse caso. Em relação ao `Drunk2Vec-SVM`, a maioria das melhorias afetou a precisão, em vez de compensar a revocação, e, portanto, o conjunto de classificadores não atingiu seu objetivo em relação ao `Drunk2Vec-SVM`.

Em conclusão, os resultados revelam que `Drunk2Ensemble` é eficaz para combinar os pontos fortes dos métodos *Drunk2Symbol* e *Drunk2Vec*, pois esse conjunto de

Tabela 6.7: Diferença entre o desempenho médio do Drunk2Ensemble e dos classificadores individuais

Base de dados	Drunk2Ensemble			Drunk2Symbol-SVM			Drunk2Vec-DNN			Drunk2Vec-SVM		
	F ₁	P	R	++ F ₁	++ P	++ R	++ F ₁	++ P	++ R	++ F ₁	++ P	++ R
DS1-drunk-mention	94,599	95,074	94,139	-4,77	-2,92	-6,62	-0,68	-1,09	-0,26	-0,15	-0,41	0,12
DS1-drunk-drinking	93,514	91,966	95,182	-4,46	-10,57	1,53	-0,44	0,85	-1,82	-0,08	-0,93	0,80
DS1-drunk-drink-now	90,611	89,930	91,385	-2,57	-9,04	3,80	-1,50	0,73	-3,65	-0,53	-1,23	0,24
DS2-keywords	90,058	89,270	90,860	-9,29	-2,56	-15,26	-4,33	-2,90	-5,60	-1,36	0,60	-3,30
DS3-drinking-ext	88,845	86,017	91,870	-5,99	-10,86	0,48	-0,41	0,08	-0,95	-0,40	-0,54	-0,23

classificadores obtêm a medida F₁ superior em todas as bases de dados. Além disso, Drunk2Ensemble equilibra as melhorias entre precisão e revocação.

Tabela 6.8: Comparativo entre Drunk2Symbol-SVM e Drunk2Ensemble

Base de dados	Drunk2Symbol-SVM			Drunk2Ensemble		
	F ₁	P	R	F ₁	P	R
DS1-drunk-mention	89,834	92,151	87,517	94,599 v	95,074 v	94,139 v
DS1-drunk-drinking	89,057	81,398	96,715	93,514 v	91,966 v	95,182
DS1-drunk-drink-now	88,037	80,892	95,182	90,611 v	89,930 v	91,385 *
DS2-keywords	80,767	86,714	75,600	90,058 v	89,270 v	90,860 v
DS3-drinking-ext	82,854	75,154	92,349	88,845 v	86,017 v	91,870

Tabela 6.9: Comparativo entre Drunk2Vec-DNN e Drunk2Ensemble

Base de dados	Drunk2Vec-DNN			Drunk2Ensemble		
	F ₁	P	R	F ₁	P	R
DS1-drunk-mention	93,919	93,979	93,880	94,599	95,074	94,139
DS1-drunk-drinking	93,072	92,815	93,362	93,514	91,966	95,182
DS1-drunk-drink-now	89,107	90,663	87,733	90,611	89,930	91,385 v
DS2-keywords	85,729	86,372	85,264	90,058 v	89,270 v	90,860 v
DS3-drinking-ext	88,437	86,102	90,923	88,845	86,017	91,870

6.6.3 Casos de falhas

Para identificar os casos de falhas dos métodos propostos, foram analisados os tweets que não foram identificados corretamente por nenhum dos seguintes classificadores: Drunk2Symbol-SVM, Drunk2Vec-DNN, Drunk2Vec-SVM e Drunk2Ensemble. A seguir, é apresentada uma análise dos casos de falhas para cada base de dados.

Para *DS1-drunk-mention*, a maioria dos tweets bêbados classificados incorretamente possuem *emoticons* relacionados ao consumo de álcool (*drinks* ou copos de *chopp*) ou imagens de bebidas alcoólicas. Por exemplo, o tweet ‘*My Memorial Day toast to my brothers-in-arms. #wine #holiday http://t.co/QmY6c5CFDY*’ contém uma imagem de uma lata de cerveja e foi classificado erroneamente como sóbrio, pois os classificadores

Tabela 6.10: Comparativo entre Drunk2Vec-SVM e Drunk2Ensemble

Base de dados	Drunk2Vec-SVM			Drunk2Ensemble		
	F ₁	P	R	F ₁	P	R
DS1-drunk-mention	94,453	94,661	94,257	94,599	95,074	94,139
DS1-drunk-drinking	93,435	91,036	95,985	93,514	91,966	95,182
DS1-drunk-drink-now	90,080	88,704	91,620	90,611	89,930 v	91,385
DS2-keywords	88,695	89,869	87,563	90,058 v	89,270	90,860 v
DS3-drinking-ext	88,440	85,474	91,635	88,845	86,017	91,870

baseiam-se apenas em *features* textuais. Para contornar o problema dos *emojicons* é possível convertê-los para símbolos alfanuméricos e inseri-los nos métodos. Alguns tweets com a expressão ‘*Drunk in love*’ (por exemplo, ‘*The amount of effort I put in when I hear drunk in love https://t.co/vOq2eOovI3 can I have him*’) foram classificados erroneamente como bêbados. Como alternativa, é possível substituir a expressão ‘*drunk in love*’ por ‘*drunkinlove*’.

Em relação a base de dados *DS1-drunk-drinking*, alguns tweets bêbados classificados incorretamente possuem termos que aparecem poucas vezes na base de dados como, por exemplo, *drunk texted*. *Drunk texted* está contido em apenas dois tweets. Por exemplo, ‘*I might’ve have drunk texted everyone I wasn’t supposed to*’. É necessário expandir a base de dados para contornar a classificação de termos pouco utilizados. Geralmente, os tweets sóbrios classificados erroneamente contêm termos relacionados ao consumo de álcool, porém em outro contexto como, por exemplo, ‘*I’ll admit whenever Fancy by IGGY comes on I scream the words like a drunk little white girl*’ que contém o termo *drunk*.

Para a base de dados *DS1-drunk-drink-now*, os classificadores enfrentaram a mesma dificuldade de *DS1-drunk-drinking*, ou seja, termos pouco utilizados. Os tweets sóbrios classificados incorretamente são menções ao consumo de álcool, porém não naquele momento. Por exemplo, ‘*I really be feeling like drinking a bottle Everytime this shit happens*’ cita a intenção de consumir álcool, mas não descreve o usuário consumindo bebidas alcoólicas no momento em que postou o tweet.

Para *DS2-keywords*, constatou-se que a maioria dos tweets bêbados classificados incorretamente são curtos, ou seja, possuem menos de cinco palavras sem contexto como, por exemplo, ‘*Never been so ha*’. Esse problema ocorre, pois, as palavras-chave foram removidas dos tweets, descaracterizando seu conteúdo. Por outro lado, os tweets sóbrios identificados erroneamente possuem referências aos termos *night* ou *bar/pub*. Por exemplo, ‘*Tap Night Tonight! Uncle Billy’s Barton Springs Pale Ale and Pedernales Lobo Negro Dark Lager*’.

Considerando *DS3-drinking-ext*, a maioria dos tweets classificados incorretamente possui a *hashtag #drunk*, visto que a mesma foi utilizada como palavra-chave. Em alguns tweets, é difícil determinar se a *hashtag #drunk* representa um estado alcoólico, visto que a mesma é utilizada em ambos os grupos e não ocorre com outras palavras relacionadas ao álcool. Por exemplo, ‘*My friend just FaceTimed his gf and she said we all look like shit #drunk*’ é anotado como bêbado, enquanto ‘*#WhitePeople love telling each other how much hey love one another when they’re #drunk*’ é anotado como sóbrio. Como alternativa, pode-se utilizar outros termos como palavras-chave para construir a base de dados. Adicionalmente, alguns tweets estão anotados erroneamente. Por exemplo, ‘*The driver is safer when the roads are dry, the roads are safer when the driver is dry. Don’t drive #drunk. #DriveSafe*’ está anotado como bêbado, porém é um tweet sóbrio. Para contornar esse problema, pode-se utilizar mais anotadores.

6.7 Experimento #4: CNN para gerar *word embeddings*

O objetivo principal deste experimento é avaliar o desempenho da rede neural CNN para gerar *word embeddings* específicas para o domínio. Nossas suposições são: *i)* algoritmos de propósito geral não possuem bom desempenho, a menos que haja um grande volume de dados para treinamento; *ii)* a eficácia das *word embeddings* pré-treinadas é limitada devido às especificidades do domínio.

6.7.1 Metodologia

Cinco variações do método *Drunk2Vec* foram criadas e executadas em todas as bases de dados. Em seguida, os tweets foram classificados utilizando os mesmos três algoritmos dos experimentos anteriores (DNN, SVM e XGBoost). Cada variação consiste em um algoritmo diferente para gerar *word embeddings* e/ou estratégia para inicializar os vetores de *embeddings*. Os resultados das variações foram comparados com o método *Drunk2Vec*. As variações são descritas a seguir:

- *SkipGram-static*: todas as bases de dados foram combinadas e foram geradas *word embeddings* específicas do domínio usando o algoritmo *Word2Vec (skip-gram)*. O *skip-gram* foi executado com os parâmetros *min_count* igual a 1, *window* igual a 10 e *negative sample* igual a 1;

- *GloVe-static*: não foram gerados *word embeddings*. Simplesmente foram utilizadas as *word embeddings* pré-treinadas do GloVe⁵ no Twitter para representar os tweets;
- *GloVe-non-static*: utilizou-se uma rede neural CNN para gerar *word embeddings*, mas em vez de inicializar aleatoriamente os vetores, fez-se uso dos pesos das *word embeddings* pré-treinadas fornecidas pelo GloVe;
- *SkipGram-non-static*: foi implementada a rede neural CNN para gerar *word embeddings*, mas em vez de inicializar aleatoriamente os vetores, empregaram-se os pesos das *word embeddings* específicas do domínio geradas usando o *Word2Vec*;
- *LSTM*: implementou-se uma rede neural LSTM, ao invés da CNN, para aprender e ajustar os pesos dos *embeddings*. *Word embeddings* foram inicializados aleatoriamente.

LSTM, *GloVe-non-static* e *SkipGram-non-static* foram treinados em cada base de dados usando dez (10) épocas com tamanho de lote igual a 64. Todos os vetores de *word embeddings* possuem dimensão igual a 100.

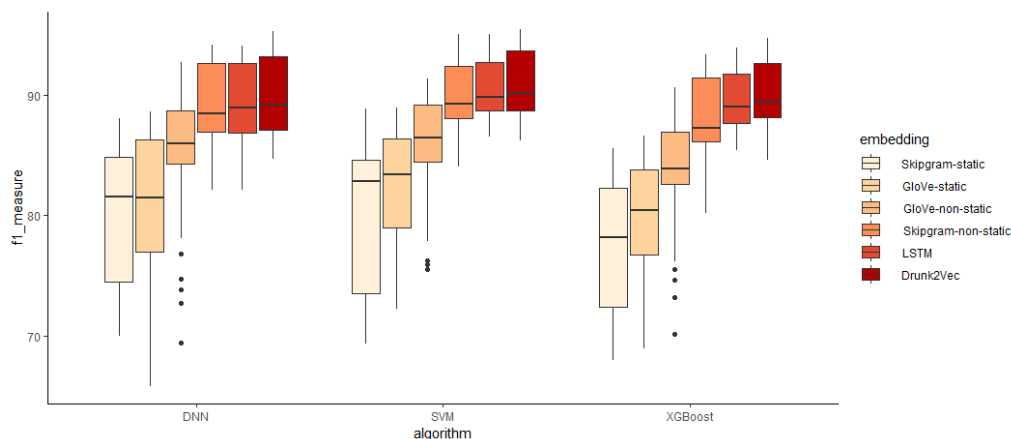
6.7.2 Resultados

A Tabela 6.12 detalha os resultados da medida F_1 para cada algoritmo de classificação (DNN, SVM, XGBoost) treinado de acordo com as variações mencionadas anteriormente, bem como *Drunk2Vec*. Os resultados englobam todas as bases de dados e são detalhados por medida F_1 mínima, Quartil inferior (Q_i), mediana, média, Quartil superior (Q_s) e medida F_1 máxima. A sumarização dos resultados é ilustrada na Figura 6.3, de modo que cada *boxplot* representa a distribuição dos resultados. Os resultados detalhados por base de dados estão disponíveis no Apêndice B.

Drunk2Vec supera todas as variações em todos os algoritmos com melhorias de até 15,68 pp. Todas as melhorias são estatisticamente significativas. Esses resultados são ligeiramente superiores aos produzidos pela LSTM, ou seja, uma variação do *Drunk2Vec* que apenas substitui a CNN por uma rede LSTM para aprender os *embeddings*. Em comparação, a mediana da medida F_1 do *Drunk2Vec* é ligeiramente superior, mas as melhorias são estatisticamente significativas (valores superiores para os quartis inferior/superior para todos os algoritmos). Em resumo, as variações baseadas em pesos inicializados ale-

⁵<http://nlp.stanford.edu/data/glove.twitter.27B.zip>

Figura 6.3: Sumarização dos resultados das variações para gerar *word embeddings*



atoriamente, mas que aprendem e ajustam os pesos relacionados aos *word embeddings* (CNN e LSTM), são as abordagens mais eficazes para aprender *drunk word embeddings*.

A variação Skipgram-static apresentou o pior desempenho para todos os classificadores, sugerindo que a combinação de todas as bases de dados não é grande o suficiente para treinar esse algoritmo adequadamente para gerar *word embeddings* específicas do domínio. Usando um modelo não estático (Skipgram-non-static), que ajusta os pesos inicializados durante o treinamento, esses resultados melhoram significativamente, mas ainda são inferiores em comparação com o *Drunk2Vec* e sua variação LSTM. De fato, os valores para as medianas e quartis inferior/superior são menores.

Os classificadores que utilizam a variação GloVe-static também apresentam um desempenho ruim, fornecendo evidências para a alegação de que *word embeddings* de propósito geral, mesmo quando treinados em um grande conjunto de dados, não representam adequadamente o vocabulário usado em tweets bêbados. Mesmo quando combinados com uma rede CNN (Glove-non-static), os resultados são inferiores.

Para analisar o valor semântico dos termos capturados por abordagens distintas para gerar *word embeddings*, foram analisadas as 10 palavras mais semelhantes aos termos *drunk*, *alcohol* e *beer*. A Tabela 6.11 descreve as palavras mais similares, considerando as *word embeddings* pré-treinadas com GloVe, *word embeddings* usando o *Word2Vec* e *word embeddings* aprendidas usando o *Drunk2Vec* (*DS1-drunk-mention*). O diferencial do *DrunkVec* é aprender as relações entre o consumo de álcool e gírias (por exemplo, *hammered*⁶, *hopping*⁷, *ratchet*⁸, *sjp*⁹, *brooke*¹⁰). *Drunk2Vec* também identifica

⁶<https://www.merriam-webster.com/dictionary/hammered>

⁷<https://www.merriam-webster.com/dictionary/hopping>

⁸<https://club937.com/what-does-the-word-ratchet-mean/>

⁹<https://www.urbandictionary.com/define.php?term=SJP>

¹⁰<https://www.urbandictionary.com/define.php?term=Brooke>

Tabela 6.11: *Embeddings* aprendidas usando CNN, *Word2Vec* e *GloVe*

Palavra	Palavras similares usando GloVe	Palavras similares usando <i>Word2Vec</i>	Palavras similares usando <i>Drunk2Vec</i>
<i>drunk</i>	<i>drunk, sober, crazy, having, silly, probably, mad, drinking, either, talking</i>	<i>drunk, lemons, inventive, annoys, dlac, grade, b2b, surround, chiacchia, cnygolf, googly, entirely</i>	<i>drunk, liquor, vodka, tequila, drinkodemayo, booze, problems, easier, shelves, hen, beverage, tipsy</i>
<i>alcohol</i>	<i>alcohol, vodka, drugs, drink, tequila, weed, consume, soda, drinks, drinking</i>	<i>alcohol, swill, huh, tooturnttuesday, became, overheardum, bedside, disliked, yeahh, seton, manner, dried</i>	<i>alcohol, tequila, crawl, booze, liquor, easier, tipsy, moon, cocktia, shelves, vodka, bud, faced, afterparty</i>
<i>beer</i>	<i>beer, wine, drink, drinks, coffee, beers, drinking, bottle, whiskey, tea</i>	<i>beer, foreal, superior, punishable, crema, fleur, craftbeertime, electrical, lemondemouth, snl, precocious, supporters, candidates</i>	<i>beer, whisky, sjp, brooke, ratchet, genesee, route, ends, bud, beers, hopping, alcoholic, hammered</i>

Tabela 6.12: Resultados das diferentes variações de *word embeddings*

Variação	DNN						SVM						XGBoost					
	Min	Qi	Mediana	Média	Qs	Max	Min	Qi	Mediana	Média	Qs	Max	Min	Qi	Mediana	Média	Qs	Max
<i>Drunk2Vec</i>	84,69	87,12	89,21	90,02	93,17	95,29	86,18	88,74	90,11	91,02	93,65	95,44	84,62	88,16	89,43	90,2	92,62	94,72
LSTM	82,13	86,88	88,92	89,21	92,67	94,08	86,52	88,67	89,82	90,61	92,69	95,05	85,39	87,67	89	89,67	91,78	93,92
Skipgram-non-static	82,11	86,97	88,44	89,18	92,63	94,18	84,06	88,09	89,25	89,91	92,4	95,09	80,16	86,17	87,27	88,16	91,44	93,33
Glove-non-static	69,42	84,31	85,99	84,94	88,7	92,7	75,56	84,42	86,48	85,73	89,2	91,37	70,11	82,57	83,87	83,7	86,95	90,62
Glove-static	65,82	76,97	81,5	81,29	86,32	88,65	72,24	78,97	83,42	82,54	86,35	88,92	68,89	76,72	80,42	79,95	83,8	86,65
Skipgram-static	69,96	74,48	81,53	79,8	84,86	88,08	69,36	73,47	82,8	80,12	84,6	88,89	67,94	72,34	78,15	77,45	82,26	85,58

a similaridade entre a palavra *drunk* e o evento *Drinko de Mayo* (*drinkodemayo*, um feriado que celebra o consumo de grandes quantidades de tequila), a relação entre *beer* e marcas de cerveja (*genesee* e *bud*) e a semelhança entre o termo *alcohol* e *cocktia* (uma versão incorreta do termo *cocktail*). Nenhum desses termos está representado nas *word embeddings* pré-treinadas do GloVe. Além disso, o GloVe relaciona o termo *sober* com seu antônimo *drunk*, o que pode comprometer a tarefa de classificação. Em relação ao *Word2Vec*, muitas palavras semelhantes não parecem estar fortemente relacionadas a esse domínio (por exemplo, *b2b* e *surround* para *drunk*; *precocious* e *supporters* para *beer*).

Concluindo, este experimento revelou que a melhoria no desempenho da classificação não está relacionada apenas à adoção de *word embeddings*, mas é fortemente influenciada pela maneira como elas são geradas. O *Drunk2Vec* superou todas as outras estratégias, usando diferentes bases de dados e algoritmos de classificação. A falta de um grande conjunto de tweets bêbados limita os benefícios de algoritmos genéricos, mesmo em modelos não estáticos. As *word embeddings* pré-treinadas não capturam as especificidades do vocabulário bêbado e enfrentam problemas de OOV.

6.8 Considerações finais

Neste capítulo foram apresentados os resultados obtidos e a metodologia experimental aplicada para validar os métodos propostos. Os resultados demonstraram melhorias estatisticamente significativas em todas as métricas de avaliação em relação aos trabalhos relacionados na área. Também é comprovada a eficácia do enriquecimento contextual externo para expandir o vocabulário, melhorando a revocação, e a importância das

word embeddings em identificar palavras similares, obtendo melhorias na precisão. O conjunto de classificadores produz os melhores resultados na medida F_1 e é eficaz para combinar ambas as abordagens de enriquecimento contextual. Por fim, foi comprovada a capacidade da rede neural CNN para aprender *word embeddings* específicas do domínio.

7 CONCLUSÃO

Neste trabalho foram propostos dois métodos para classificar tweets escritos sob influência do álcool. *Drunk2Symbol* explora o enriquecimento contextual externo para extrair conhecimento e fornecer contexto aos tweets, enquanto o *Drunk2Vec* faz uso do aprendizado profundo e da semântica distribucional para capturar as relações sintáticas e semânticas usadas nos tweets bêbados. Os métodos também são compostos de etapas para tratar erros, integrar *features* e lidar com a alta dimensionalidade. Em conjunto com algoritmos de classificação, esses métodos apresentam alta assertividade para filtrar conteúdos relacionados ao consumo de álcool nas redes sociais. Diferentemente dos trabalhos relacionados analisados, este trabalho aplica *word embeddings*, enriquecimento contextual externo e um conjunto de classificadores para identificar tweets bêbados, não se limitando as *features* textuais.

Este trabalho também descreve a construção de duas bases de dados públicas relacionadas ao consumo de álcool. As bases de dados, denominadas *DS2-keywords* e *DS3-drinking-ext*, são importantes contribuições deste trabalho, uma vez que existem poucas bases de dados disponíveis, dificultando a realização de experimentos nesta área. As bases de dados podem ser utilizadas por outros trabalhos para analisar e classificar o consumo de álcool nas redes sociais.

A análise exploratória fornece informações relevantes relacionadas ao consumo de álcool, apresentando mudanças na forma com que as emoções são expressas por usuários possivelmente alcoolizados. Essas mudanças podem explicar a motivação para ingerir bebidas alcoólicas (por exemplo, redução da tristeza e raiva, aumento da alegria) e a acidentes (por exemplo, redução do medo). A análise exploratória também identificou a relação entre estudantes e o consumo de álcool.

Os experimentos avaliaram os métodos utilizando diferentes algoritmos de classificação e cinco bases de dados, que abordam diferentes comportamentos do consumo de álcool no Twitter. Os resultados demonstram que cada método desempenha um papel diferente na classificação de textos bêbados. *Drunk2Symbol* melhora a revocação, lidando com a esparcialidade e os ruídos dos tweets, enquanto *Drunk2Vec* produz melhores resultados na precisão, superando as idiossincrasias da linguagem utilizada em tweets bêbados. Um conjunto de classificadores foi implementado para combinar os pontos fortes dos métodos *Drunk2Symbol* e *Drunk2Vec* e equilibrar as melhorias entre precisão e revocação. Em relação ao *baseline*, foram obtidas melhorias estatisticamente significativas na medida

F_1 , variando entre 5,9 pp a 11,6 pp. Também foram obtidas melhorias de até 13,7 pp na revocação e 12,2 pp na precisão.

A avaliação detalhada do papel de cada método também é uma contribuição deste trabalho. Essa avaliação confirmou a importância das *features semânticas* para o *Drunk2Symbol* e do papel complementar de ambas as abordagens de enriquecimento externo (NLU e Web Semântica). Também foi confirmada a suposição de que um conjunto de classificadores é eficaz para combinar ambas as abordagens de enriquecimento contextual. Além disso, as descobertas deste trabalho sugerem que a rede neural CNN captura diferentes nuances da linguagem empregada em tweets bêbados, como gírias, erros ortográficos e eventos relacionados ao álcool. Essa avaliação é valiosa para o desenvolvimento de novas abordagens para a classificação de tweets, não se limitando a assuntos relacionados ao consumo de álcool.

Inicialmente, a proposta desta dissertação foi publicada no Workshop de Teses e Dissertações em Banco de Dados (WTDBD) (GRZEÇA; BECKER; GALANTE, 2018a). Um artigo contendo o método *Drunk2Symbol* e a análise exploratória dos dados foi aceito e publicado como artigo regular na *International Conference on Web Intelligence 2018* (GRZEÇA; BECKER; GALANTE, 2018b). Por fim, um artigo contendo os métodos *Drunk2Symbol* e *Drunk2Vec*, descritos nesta dissertação, foi escrito e submetido a revista *Information Processing & Management* (IP&M).

Para trabalhos futuros, pretende-se estender os métodos propostos a outros trabalhos que usam dados de redes sociais para analisar o comportamento humano como, por exemplo, o consumo de drogas ilícitas e a identificação de usuários com depressão. O estudo desta dissertação foi limitado a tweets em inglês, mas pode ser estendido para outros idiomas, em especial o idioma português. Também se planeja identificar textos bêbados a partir de qualquer formato textual, não se restringindo a tweets, pois existem outras redes sociais (por exemplo, *Facebook*, *Instagram*) que são utilizadas com frequência por jovens e possuem textos vinculados ao consumo de álcool. Outras oportunidades são aumentar os tamanhos das bases de dados e realizar uma análise detalhada dos falsos positivos e falsos negativos para cada método, identificando casos de falha e possíveis melhorias. A análise exploratória pode ser estendida para incorporar outras visualizações como, por exemplo, análise de tópicos e sumarização de eventos relacionados ao consumo de álcool.

REFERÊNCIAS

AGARWAL, B. et al. A deep network model for paraphrase detection in short text messages. **Inf. Process. Manage.**, v. 54, n. 6, p. 922–937, 2018. Available from Internet: <<https://doi.org/10.1016/j.ipm.2018.06.005>>.

AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. In: **Mining Text Data**. [s.n.], 2012. p. 163–222. Available from Internet: <https://doi.org/10.1007/978-1-4614-3223-4_6>.

AN, J.; WEBER, I. #greysanatomy vs. #yankees: Demographics and hashtag use on twitter. In: **Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016**. [s.n.], 2016. p. 523–526. Available from Internet: <<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13021>>.

APHINYANAPHONGS, Y. et al. Text classification for automatic detection of alcohol use-related tweets: A feasibility study. In: **Proceedings of the 15th IEEE International Conference on Information Reuse and Integration, IRI 2014, Redwood City, CA, USA, August 13-15, 2014**. [s.n.], 2014. p. 93–97. Available from Internet: <<https://doi.org/10.1109/IRI.2014.7051877>>.

BADJATIYA, P. et al. Deep learning for hate speech detection in tweets. In: **Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017**. [s.n.], 2017. p. 759–760. Available from Internet: <<https://doi.org/10.1145/3041021.3054223>>.

BENTON, A.; ARORA, R.; DREDZE, M. Learning multiview embeddings of twitter users. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers**. [s.n.], 2016. p. 14–19. Available from Internet: <<https://www.aclweb.org/anthology/P16-2003/>>.

BOEHMKE, B.; GREENWELL, B. M. **Hands-On Machine Learning with R**. [S.l.]: CRC Press, 2019.

BORRILL, J. A.; ROSEN, B. K.; SUMMERFIELD, A. B. The influence of alcohol on judgement of facial expression of emotion. **The British journal of medical psychology**, v. 60, p. 71–77, 1987.

CAVAZOS-REHG, P. A. et al. “hey everyone, i’m drunk.” an evaluation of drinking-related twitter chatter. **Journal of studies on alcohol and drugs**, Rutgers University, v. 76, n. 4, p. 635–643, 2015.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **ACM. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.], 2016. p. 785–794.

CHOLLET, F.; ALLAIRE, J. J. **Deep Learning with R**. [S.l.]: Manning Publications Company, 2018.

CULOTTA, A. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. **Language Resources and Evaluation**, v. 47, n. 1, p. 217–238, 2013. Available from Internet: <<https://doi.org/10.1007/s10579-012-9185-0>>.

CURTIS, B. et al. Can twitter be used to predict county excessive alcohol consumption rates? **PloS one**, Public Library of Science, v. 13, n. 4, p. e0194290, 2018.

FALOTICO, R.; QUATTO, P. Fleiss' kappa statistic without paradoxes. **Quality & Quantity**, Springer, v. 49, n. 2, p. 463–470, 2015.

FILHO, R. M.; CARVALHO, A. I.; PAPP, G. L. Inferência de sexo e idade de usuários no twitter. 2014.

GHAFARIAN, S. H.; YAZDI, H. S. Identifying crisis-related informative tweets using learning on distributions. **Information Processing & Management**, Elsevier, v. 57, n. 2, p. 102145, 2020.

GOLDBERG, Y. A primer on neural network models for natural language processing. **J. Artif. Intell. Res.**, v. 57, p. 345–420, 2016. Available from Internet: <<https://doi.org/10.1613/jair.4992>>.

GOLDBERG, Y. Neural network methods for natural language processing. **Synthesis Lectures on Human Language Technologies**, Morgan & Claypool Publishers LLC, v. 10, n. 1, p. 1–309, apr 2017. Available from Internet: <<https://doi.org/10.2200/s00762ed1v01y201703hlt037>>.

GOODFELLOW, I. et al. **Deep learning**. [S.l.]: MIT press Cambridge, 2016.

GRZEÇA, M. A.; BECKER, K.; GALANTE, R. Effective method for detecting drunk texting. In: **SBBD Companion**. [S.l.: s.n.], 2018. p. 60–66.

GRZEÇA, M. A.; BECKER, K.; GALANTE, R. Improving the classification of drunk texting in tweets using semantic enrichment. In: **2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018, Santiago, Chile, December 3-6, 2018**. [s.n.], 2018. p. 190–197. Available from Internet: <<https://doi.org/10.1109/WI.2018.00-90>>.

GUPTA, I.; JOSHI, N. Tweet normalization: A knowledge based approach. In: **IEEE. 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS)**. [S.l.], 2017. p. 157–162.

HARB, J. G. D.; BECKER, K. Comparing emotional reactions to terrorism events on twitter. In: **Big Social Data and Urban Computing - First Workshop, BiDU@VLDB 2018, Rio de Janeiro, Brazil, August 31, 2018, Revised Selected Papers**. [s.n.], 2018. p. 107–122. Available from Internet: <https://doi.org/10.1007/978-3-030-11238-7_7>.

HASAN, M.; AGU, E.; RUNDENSTEINER, E. Using hashtags as labels for supervised learning of emotions in twitter messages. In: **ACM SIGKDD Workshop on Health Informatics, New York, USA**. [S.l.: s.n.], 2014.

HEIMISDOTTIR, J. et al. The social context of drunkenness in mid-adolescence. **Scandinavian journal of public health**, Sage Publications Sage UK: London, England, v. 38, n. 3, p. 291–298, 2010.

- HOSSAIN, N. et al. Precise localization of homes and activities: Detecting drinking-while-tweeting patterns in communities. In: **Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016**. [s.n.], 2016. p. 587–590. Available from Internet: <<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13118>>.
- HU, H. et al. Deep self-taught learning for detecting drug abuse risk behavior in tweets. In: **Computational Data and Social Networks - 7th International Conference, CSoNet 2018, Shanghai, China, December 18-20, 2018, Proceedings**. [s.n.], 2018. p. 330–342. Available from Internet: <https://doi.org/10.1007/978-3-030-04648-4_28>.
- JAUCH, A.; JAEHNE, P.; SUENDERMANN, D. Using text classification to detect alcohol intoxication in speech. In: **Proceedings of the 7th Workshop on Emotion and Computing at the 36th German Conference on Artificial Intelligence**. [S.l.: s.n.], 2013.
- JERNIGAN, D. et al. Alcohol marketing and youth alcohol consumption: a systematic review of longitudinal studies published since 2008. **Addiction**, Wiley Online Library, v. 112, p. 7–20, 2017.
- JOHNSON, T.; SHAPIRO, R.; TOURANGEAU, R. National survey of american attitudes on substance abuse xvi: Teens and parents. **The National Center on Addiction and Substance Abuse**, v. 2011, 2011.
- JOSHI, A. et al. A computational approach to automatic prediction of drunk-texting. In: **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers**. [s.n.], 2015. p. 604–608. Available from Internet: <<https://www.aclweb.org/anthology/P15-2100/>>.
- KHATUA, A.; KHATUA, A.; CAMBRIA, E. A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks. **Inf. Process. Manage.**, v. 56, n. 1, p. 247–257, 2019. Available from Internet: <<https://doi.org/10.1016/j.ipm.2018.10.010>>.
- KIM, Y. Convolutional neural networks for sentence classification. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL**. [s.n.], 2014. p. 1746–1751. Available from Internet: <<https://www.aclweb.org/anthology/D14-1181/>>.
- KNOX, J. et al. Using social network analysis to examine alcohol use among adults: A systematic review. **PloS one**, Public Library of Science, v. 14, n. 8, 2019.
- KORDE, V.; MAHENDER, C. N. Text classification and classifiers: A survey. **International Journal of Artificial Intelligence & Applications**, Academy & Industry Research Collaboration Center (AIRCC), v. 3, n. 2, p. 85, 2012.
- LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, JSTOR, p. 159–174, 1977.

LI, Q. et al. Tweetsift: Tweet topic classification based on entity knowledge base and topic enhanced word embedding. In: **ACM. Proc. of the 25th ACM International on Conference on Information and Knowledge Management**. [S.l.], 2016. p. 2429–2432.

MAITY, S. K. et al. Understanding psycholinguistic behavior of predominant drunk texters in social media. In: **2018 IEEE Symposium on Computers and Communications, ISCC 2018, Natal, Brazil, June 25-28, 2018**. [s.n.], 2018. p. 1096–1101. Available from Internet: <<https://doi.org/10.1109/ISCC.2018.8538637>>.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: **Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States**. [s.n.], 2013. p. 3111–3119. Available from Internet: <<http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality>>.

MIZZARO, S. et al. Short text categorization exploiting contextual enrichment and external knowledge. In: **ACM. Proceedings of the first international workshop on Social media retrieval and analysis**. [S.l.], 2014. p. 57–62.

MYSLÍN, M. et al. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. **Journal of medical Internet research**, JMIR Publications Inc., v. 15, n. 8, 2013.

ORGANIZATION, W. H. **Global status report on alcohol and health 2018**. [S.l.]: World Health Organization, 2019.

PARROTT, D. J.; ECKHARDT, C. I. Effects of alcohol on human aggression. **Current Opinion in Psychology**, Elsevier, v. 19, p. 1–5, 2018.

PENG, Z.; HU, Q.; DANG, J. Multi-kernel SVM based depression recognition using social media data. **Int. J. Machine Learning & Cybernetics**, v. 10, n. 1, p. 43–57, 2019. Available from Internet: <<https://doi.org/10.1007/s13042-017-0697-1>>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL**. [s.n.], 2014. p. 1532–1543. Available from Internet: <<https://www.aclweb.org/anthology/D14-1162/>>.

ROLLER, R.; THOMAS, P.; SCHMEIER, S. Football and beer - a social media analysis on twitter in context of the FIFA football world cup 2018. **CoRR**, abs/1811.03809, 2018. Available from Internet: <<http://arxiv.org/abs/1811.03809>>.

ROMERO, S.; BECKER, K. A framework for event classification in tweets based on hybrid semantic enrichment. **Expert Systems with Applications**, v. 118, p. 522 – 538, 2019. ISSN 0957-4174. Available from Internet: <<http://www.sciencedirect.com/science/article/pii/S095741741830678X>>.

SCHIEL, F.; HEINRICH, C.; BARFÜSSER, S. Alcohol language corpus: the first public corpus of alcoholized german speech. **Language resources and evaluation**, Springer, v. 46, n. 3, p. 503–521, 2012.

- SCHULZ, A.; GUCKELSBERGER, C.; JANSSEN, F. Semantic abstraction for generalization of tweet classification: An evaluation of incident-related tweets. **Semantic Web**, v. 8, n. 3, p. 353–372, 2017. Available from Internet: <<https://doi.org/10.3233/SW-150188>>.
- SEO, Y.-W. **InfoGain**. 2018. Available from Internet: <<http://www.cs.cmu.edu/~youngwoo/projects/textminer/textminer-docs/doc/textminer/featureselection/InfoGain.html>>.
- STEIN, R. A.; JQUES, P. A.; VALIATI, J. F. An analysis of hierarchical text classification using word embeddings. **Inf. Sci.**, v. 471, p. 216–232, 2019. Available from Internet: <<https://doi.org/10.1016/j.ins.2018.09.001>>.
- TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: A review. **Data classification: Algorithms and applications**, CRC Press, p. 37, 2014.
- WANG, P. et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. **Neurocomputing**, v. 174, p. 806–814, 2016. Available from Internet: <<https://doi.org/10.1016/j.neucom.2015.09.096>>.
- WEST, J. H. et al. Temporal variability of problem drinking on twitter. **Open Journal of Preventive Medicine**, Scientific Research Publishing, v. 2, n. 01, p. 43, 2012.
- WITTEN, I. H. et al. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016.
- YATES, A.; COHAN, A.; GOHARIAN, N. Depression and self-harm risk assessment in online forums. In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017**. [s.n.], 2017. p. 2968–2978. Available from Internet: <<https://www.aclweb.org/anthology/D17-1322/>>.
- ZHANG, D.; HE, D. Can word embedding help term mismatch problem? - A result analysis on clinical retrieval tasks. In: **Transforming Digital Worlds - 13th International Conference, iConference 2018, Sheffield, UK, March 25-28, 2018, Proceedings**. [s.n.], 2018. p. 402–408. Available from Internet: <https://doi.org/10.1007/978-3-319-78105-1_44>.
- ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis: A survey. **Wiley Interdiscip. Rev. Data Min. Knowl. Discov.**, v. 8, n. 4, 2018. Available from Internet: <<https://doi.org/10.1002/widm.1253>>.

APÊNDICE A — ÍNDICES DE CONCORDÂNCIA DO COEFICIENTE KAPPATabela A.1: Índices de concordância do coeficiente *kappa*

<i>Kappa Statistic</i>	<i>Strength of Agreement</i>
< 0	<i>Poor</i>
0,00 - 0,20	<i>Slight</i>
0,21 - 0,40	<i>Fair</i>
0,41 - 0,60	<i>Moderate</i>
0,61 - 0,80	<i>Substantial</i>
0,81 - 1,00	<i>Almost Perfect</i>

Fonte: (LANDIS; KOCH, 1977)

APÊNDICE B — VARIAÇÕES DAS *WORD EMBEDDINGS* POR BASE DE DADOS

As Tabelas B.1, B.2, B.3 exibem os resultados do Experimento #4 detalhado por base de dados. Os valores representam a média da medida F_1 para cada variação e base de dados. O símbolo (*) representa que o método *Drunk2Vec* é estatisticamente superior, enquanto que o símbolo (v), em negrito, indica que as melhorias obtidas pela variação proposta são estatisticamente significativas.

Tabela B.1: DNN

Variação	DS1-Q1	DS1-Q2	DS1-Q3	DS2	DS3
Drunk2Vec	93,92	93,07	89,11	85,73	88,44
Glove-static	87,17 *	86,67 *	73,43 *	77,44 *	81,75 *
Glove-non-static	90,64 *	88,21 *	75,72 *	85,53	84,61 *
Skipgram-static	84,08 *	86,42 *	74,32 *	72,73 *	81,43 *
Skipgram-non-static	93,14	92,06 *	85,34 *	87,11 v	88,26
LSTM	93,30 *	92,38 *	88,92	83,73 *	87,73

Tabela B.2: SVM

Variação	DS1-Q1	DS1-Q2	DS1-Q3	DS2	DS3
Drunk2Vec	94,45	93,43	90,08	88,70	88,44
Glove-static	87,22 *	85,98 *	88,92 *	78,85 *	82,37 *
Glove-non-static	90,20 *	88,19 *	77,56 *	86,43 *	84,82 *
Skipgram-static	84,39 *	85,66 *	79,15 *	73,06 *	82,37 *
Skipgram-non-static	93,74	91,26 *	73,36 *	88,60	88,56
LSTM	94,29	92,15 *	86,35 *	88,87	89,09

Tabela B.3: XGBoost

Variação	DS1-Q1	DS1-Q2	DS1-Q3	DS2	DS3
Drunk2Vec	93,32	92,50	88,83	88,23	88,02
Glove-static	85,70 *	83,34 *	73,35 *	76,66 *	80,18 *
Glove-non-static	89,10 *	86,43 *	75,73 *	83,82 *	82,35 *
Skipgram-static	82,02 *	83,24 *	72,57 *	71,37 *	78,20 *
Skipgram-non-static	92,70 *	90,90 *	83,92 *	86,63 *	87,00 *
LSTM	93,44	91,46 *	87,89	88,36	87,75