Avaliação de métodos de desofuscação de palavrões

Evaluation of profanity deobfuscation methods

Gustavo Laboreiro
Faculdade de Engenharia da Universidade do Porto - DEI - LIACC
gustavo.laboreiro@fe.up.pt

Eugénio Oliveira
Faculdade de Engenharia da Universidade do Porto - DEI - LIACC
eco@fe.up.pt

Resumo

Os palavrões são uma forma de expressão notável pela sua intensidade. Quando uma pessoa recorre a este tipo de expressão emite uma opinião ou avaliação espontânea e crua, que muitas vezes é suprimida em nome dos "bons costumes" e sensibilidades. Acontece que esta forma de expressão tem também valor aquando de certas análises de opinião e de sentimentos, que são operações já feitas de forma rotineira nas redes sociais. Daí que neste trabalho procuramos avaliar os métodos que permitem recuperar e reconhecer estas formas de expressão que foram disfarçadas através de ofuscação, muitas vezes como forma de evasão à censura automática.

Keywords

palavrões, obscenidades, conteúdo gerado pelo utilizador.

Abstract

Cursing is a form of expression that is noted by its intensity. When someone uses this form of expression they are emitting a spontaneous and raw form of opinion, usually suppressed for the "mild ways" and sensitive people. As it happens, this sort of expression is also valuable when doing some sort of opinion mining and sentiment analysis, now a routine task across the social networks. Therefore in this work we try to evaluate the methods that allow the recovery of this forms of expression, disguised through obfuscation methods, often as a way to escape automatic censorship.

Keywords

profanity, cursing, user-generated content.

1 Introdução

A *Web 2.0* trouxe um novo paradigma à Internet, ao encorajar os utilizadores a partilhar conteúdo de sua própria autoria, a deixar os seus juízos e avaliações, e até a interagir com os autores e outros comentadores através de conversas e trocas de impressões. Mas os responsáveis pela plataforma têm de zelar belo seu bom funcionamento, e garantir que alguns utilizadores mal-comportados não afastam o público-alvo. Esta situação é mais significativa quando o conteúdo primário é da responsabilidade dos donos da plataforma (como é o caso de jornais *on-line*).

Um dos abusos que se pretende evitar é o uso de discurso insultuoso, e em particular o recurso aos palavrões. Tem sido este o objetivo dos estudos sobre o tema: procurar palavras presentes num léxico "proibido". Mas há muito mais que os palavrões podem dizer-nos a diversos níveis, e que tem sido ignorado.

O nosso objetivo é procurar uma forma de permitir que os palavrões sejam usados como um recurso adicional na análise automática de utilizadores ou mensagens. Mas para tal é necessário ser capaz de os identificar e reconhecer no meio do texto ruidoso, mesmo que estejam "disfarçados". Neste trabalho iremos analisar a principais formas de abordar este desafio, os seus problemas e os seus resultados.

Iremos explorar primeiro os palavrões e os seus usos, assim como os trabalhos científicos que os abordaram, a fim de estabelecer o contexto em que o presente trabalho se insere. De seguida será apresentada a coleção anotada que serviu de base à análise que é feita, a coleção "SAPO Desporto", que se encontra disponível on-line. A metodologia da experiência é apresentada na Secção 4, onde detalhamos os processos que foram considerados relevantes para a correta identificação dos palavrões, e como podem afetar os resultados. Na Secção 5 detalhamos os nossos testes, cujos re-

sultados acompanhamos dos nossos comentários e análise. Concluímos o nosso trabalho sumariando as nossas conclusões e apresentamos as nossas propostas para trabalho futuro.

1.1 O que são os palavrões, e qual a sua utilidade?

Neste trabalho definimos palavrões como palavras socialmente convencionadas indecentes que são usadas com intenções ofensivas ou vulgares. A origem, a disseminação, a interpretação e o uso de palavrões podem ser estudados em várias áreas, como a psicologia, a linguística, a antropologia e a sociologia.

Sabemos que alguns estados emocionais podem ser expressos de forma adequada apenas através do uso de palavrões (Jay e Janschewitz, 2007), em particular a frustração, a fúria e a surpresa (Jay, 2009) ou a tristeza e a fúria (Wang et al., 2014). Daí que será expectável que os palavrões sejam significantes na análise automática de sentimentos ou opiniões (Constant et al., 2009).

O uso de palavrões tende também a variar em função de diversos fatores associados ao orador, nomeadamente com a sua idade, sexo, identidade de grupo, fatores de personalidade e classe social (Thelwall, 2008; Jay, 2009), e como tal pode ajudar na elaboração de perfis de utilizadores.

1.2 Quão frequente é o uso de palavrões?

É claro que a resposta a esta pergunta depende de muitos fatores — quem, onde, em que contexto, etc.. Mas a título inicial podemos começar pelo estudo de Mehl e Pennebaker, que analisaram 4 dias de gravações contendo as conversas e interações de 52 estudantes universitários (Mehl e Pennebaker, 2003), resultando na estimativa de que 0,5% das palavras proferidas eram palavrões. Este estudo estabelece uma base de comparação que podemos usar na análise dos três estudos seguintes, que abordam o uso de palavrões em três comunidades on-line diferentes.

1.2.1 No MySpace

Em 2006, Thelwall fez um estudo relativo à rede social MySpace, onde comparou o uso de palavrões entre 9376 perfis, divididos entre utilizadores dos Estados Unidos da América e do Reino Unido (Thelwall, 2008). O autor encontrou palavrões nas *homepages* da maioria de jovens (considerando como jovens todos os utilizadores com idade igual ou inferior aos 19 anos), sendo que

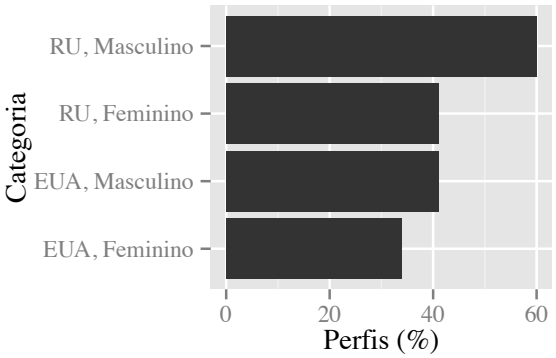

Figura 1: Proporção dos 9376 perfis do MySpace analisados que continham profanidade (média, grave ou muito grave), agrupados por sexo e nacionalidade.

a nível geral quase 40% dos perfis apresentavam este vocabulário.

A Figura 1 ilustra a prevalência de palavrões entre os dois sexos e as duas nacionalidades em estudo, mostrando que esta linguagem é mais prevalecente entre os utilizadores do sexo masculino e na comunidade do Reino Unido. A grande maioria dos palavrões encontrados foram classificados como sendo "fortes", como mostra a Figura 2. Os palavrões moderados tinham maior presença no Reino Unido (onde é 2 a 3 vezes mais comum), enquanto que o recurso a linguagem "muito forte", apesar de comparativamente rara, se encontrava também muito mais presente na comunidade britânica.

O estudo de Thelwall focou-se mais nos palavrões fortes e muito fortes aquando do estudo da frequência das palavras. Com base neste vocabulário mais restrito aponta como rácio máximo de palavrões por palavra empregue de 5% de palavrões para um utilizador do Reino Unido, ou seja, observaram um britânico empregava um palavrão a cada 20 palavras na sua *homepage*. Para um norte-americano o rácio máximo que foi visto foi de 11%, o que é significativamente superior. No entanto, estes valores extremos não são muito representativos, como se pode perceber pela Tabela 1, que compila os valores médios.

1.2.2 No Twitter

Mais recentemente, em 2014, Wang et al. estudaram 14 milhões de utilizadores do Twitter, que entre si englobavam 51 milhões de mensagens em inglês (Wang et al., 2014). Nelas, 7,73% continham pelo menos um palavrão (1 em 13 *tweets*), sendo que estes eram observados com uma frequência de 1,15% referente ao total de pala-

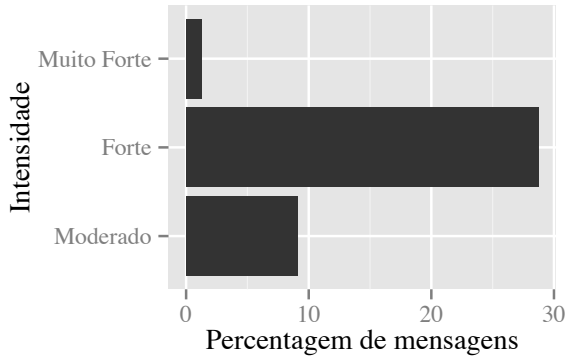

Figura 2: Classificação do nível de intensidade dos palavrões encontrados nas mensagens analisadas do MySpace.

País	Classe	Percentagem
RU	Geral	0,2
	Masculino	0,23
	Feminino	0,15
EUA	Geral	0,3
	Masculino	0,3
	Feminino	0,2

Tabela 1: Percentagem média de palavras consideradas como palavrões fortes ou muito fortes, nos perfis amostrados do MySpace.

vras escritas — o que faz uma média de um palavrão por cada 87 palavras observadas. Dado que 1% das palavras tipicamente usadas em conversas orais nessa língua são pronomes pessoais na primeira pessoa (e.g. "we", "us", "our") (Mehl e Pennebaker, 2003), e os linguistas não os consideram como termos raros, podemos afirmar que este tipo de linguagem tem presença significativa no Twitter.

Os autores referem também que os palavrões são associados principalmente com emoções negativas. As mensagens com palavrões exprimiam mais emoções como tristeza ou fúria, e muito poucas abordavam o amor, como transcrito na Tabela 2. Por outro lado, como apresentado na Tabela 3, mais de um em cada cinco *tweets* zangados contém palavrões, estando comparativamente muito ausentes das mensagens positivas.

	Tristeza	Fúria	Amor
Com palavrões	21,83	16,79	6,59
Sem palavrões	11,31	4,50	nd

Tabela 2: Percentagem de *tweets* contendo ou não palavrões que exprimem as seguintes emoções.

Emoção	Percentagem
Fúria	23,82
Tristeza	13,93
Amor	4,16
Agradecimento	3,26
Alegria	2,50

Tabela 3: Percentagem de *tweets* exprimindo diversas emoções que contêm palavrões.

1.2.3 No Yahoo! Buzz

Sood, Antin e Churchill publicaram vários trabalhos relacionados com o estudo da linguagem e comunidades on-line. Baseado na análise de 6500 mensagens extraídas do Yahoo! Buzz (Sood, Churchill e Antin, 2011), uma comunidade social centrada na submissão e comentário de notícias on-line, derivada do Digg. Este estudo tem a particularidade de ter recorrido à plataforma de *croudsourcing* da Amazon, o Mechanical Turk, para analisar as mensagens.

Os restantes estudos aqui citados limitaram-se a procurar por palavras presentes num léxico, assim como algumas variantes, o que lhes permitiu abordar grandes quantidades de dados rapidamente. Contudo, os palavrões ou suas variantes que não foram considerados no léxico (e há sempre vários — abordaremos também a questão da ofuscação mais à frente) não foram contabilizados. Logo, há um problema de *cobertura*, ou seja, há a forte possibilidade de haver *falsos negativos* (por vezes chamados de "erros do Tipo II"). Os números referidos tendem por isso a ser mais comedidos e a pecar por diferença.

Sood e seus colaboradores recorreram a anotadores que classificam todas as mensagens, revelando o número de palavrões existentes na coleção (salvo a discordância de interpretação entre os anotadores). No entanto, a consequência é que as coleções tendem a ser de dimensão menor, e consequentemente, menos representativas da população.

Os resultados desta análise (Sood, Antin e Churchill, 2012a) indicam que das 6354 mensagens (146 não obtiveram consenso entre os anotadores), 9,28% continham 1 ou mais palavrões. Infelizmente não foram apresentados valores para a percentagem de palavras que eram palavrões — possivelmente porque a anotação foi feita ao nível da mensagem. A Figura 3 apresenta os resultados referentes a 10 secções de notícias, mostrando que este tipo de vocabulário era mais prevalente na secção de política e menos usado nos comentários desportivos.

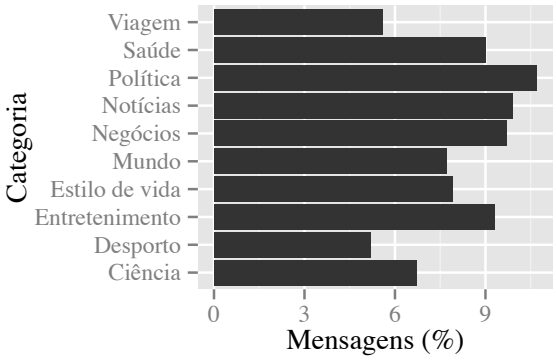

Figura 3: Proporção das mensagens nas diversas secções no Yahoo! Buzz que continham profanidade.

Por esta altura podemos concluir duas coisas: os palavrões têm uma presença significativa nas interações on-line, e que eles variam de acordo com uma série de fatores — como o país, o sexo, a idade, os sentimentos e os assuntos em discussão. Esta variação não se reflete apenas na frequência, mas também nas palavras que são empregues.

1.3 Quantos palavrões existem?

Pode afirmar-se à partida que é impossível compilar uma coleção que contenha todos os palavrões, e por vários motivos.

Como tudo o que é de cultura tradicionalmente oral, os palavrões evoluíram de forma divergente e apresentam múltiplas variações. Basta consultar uma compilação de expressões carroceiras (Almeida, 2014) para perceber que o contexto e a interpretação são muito importantes na identificação e compreensão de muitas palavras e expressões. Muitas palavras inócuas podem ser combinadas numa expressão insultuosa — e, por vezes, via este processo uma palavra pode começar a adquirir uma conotação mais "mal vista".

Focando-nos mais em palavras individuais, os palavrões são encontrados raramente na forma escrita em publicações de referência (como dicionários, livros e imprensa); e a pouca exposição à sua grafia correta favorece a confusão sobre como se escrevem. Assim, torna-se mais provável a uma pessoa escrever mal um palavrão de forma acidental. Por exemplo, escreve-se "murcão" ou "morcão"? A consequência desta dúvida comum é a facilidade de se ir encontrando, ao longo do tempo, o mesmo palavrão escrito de várias formas diferentes, tornando indistinta a forma correta. A este problema devemos ainda acrescentar as variações regionais e culturais que são transversais a todo o português. Desta forma cada palavrão acumula um número variado de formas de ser escrito.

Aliadas também às questões etnográficas há ainda a questão que temos ignorado até aqui de que a nossa definição de palavrão está sujeita a *interpretação*. Por exemplo, no Brasil "veado" é um termo insultuoso associado à homossexualidade ou efeminação, e o mesmo acontece com muitos termos que lhe são associados, como "Bambi" ou o número 24 — o número do veado no Jogo do Bicho[1]. Em oposição, a cultura Portuguesa tem o veado como um símbolo antigo de masculinidade. Outro exemplo de interpretação diferente de é a palavra "cu", que pode ser considerada indelicada ou mal vista por algumas pessoas, enquanto que outras a usam sem qualquer hesitação, independentemente do contexto social em que se encontram.

Sood, Antin e Churchill abordam de certa forma esta questão, referindo que o contexto dita qual o sentido do uso de uma palavra, e como pode ditar quando uma palavra é ou não aceitável (Sood, Antin e Churchill, 2012a). Por exemplo, "cornudo" é um termo aceitável quando o tema são animais, como discussões de zoologia, biologia ou pecuária; mas noutros pode visto como insultuoso. Nem sempre é fácil deduzir o contexto em que um termo é empregue, e até lá não podemos assumir que um sistema automático de deteção de palavrões baseados em listas consiga ter sucesso pleno.

Devemos também recordar que os palavrões estão sujeitos às mesmas variações que afetam todas as outras palavras, tais como género, número, diminutivos, aumentativos, conjunções, etc.. Existe também a possibilidade de criar novas palavrões através de aglutinação (Thelwall, 2008), embora seja uma ocorrência mais comum no inglês. Contabilizando todos estes fatores, mesmo começando com uma lista de palavrões modesta, facilmente se expande até às muitas centenas de elementos se tentarmos contabilizar as variantes possíveis.

Se não é fácil reconhecer de forma automática um palavrão dadas todas as formas como ele pode ser escrito quando o autor *quer* que ele seja percebido, o problema fica muito mais complicado quando o autor *não quer* ser óbvio, e ergue propositadamente uma barreira à sua compreensão. Iremos de seguida abordar a questão da ofuscação de palavrões, que traz complicações acrescidas a todos os processos que lidam com os palavrões.

[1] Uma espécie de lotaria popular no Brasil, que é ilegal mas tolerada pelas autoridades.

1.4 O que é a ofuscação?

Chamamos ofuscação ao desvio ou alteração propositada da grafia de uma palavra da sua versão canónica. Podemos enunciar vários motivos que podem levar o autor a recorrer à ofuscação na elaboração do seu texto.

Estilístico Escrita que invoca uma determinada pronúncia ou forma de falar (por exemplo: "Bibó Puorto!");

Diferenciador Escrever de forma diferente a fim de mostrar uma certa identidade de grupo. Por exemplo, o chamado "leet speak" (1337);

Crítica A alteração mais ou menos grave, normalmente de um nome, de forma a acarretar algum juízo de valor sobre o mesmo. Por exemplo, "Micro$oft" para indicar que considera que a empresa liga apenas a questões financeiras, ou "Pinócrates" para sublinhar que José Sócrates um mentiroso na sua opinião;

Auto-censura Para quando uma pessoa não quer ser perfeitamente explícita, como por exemplo "m*rda", ou quando se pretende contornar mecanismos de controlo baseados no conteúdo das mensagens, como filtros anti-spam (e.g. "V!@gra");

Abreviação Quando não quer escrever todos os carateres, como em "FdP".

Dado que as intenções dos autores nem sempre são aparentes, não temos sempre conhecimento se dada palavra foi escrita de forma diferente propositadamente ou de forma acidental. Como do ponto de vista do nosso trabalho essa distinção não é significativa, iremos considerar todos os desvios gráficos na escrita de palavrões como tentativas de ofuscação.

1.5 Porquê apostar na desofuscação?

O principal foco dos trabalhos que lidam com palavrões é saber se determinada mensagem tem ou não palavrões, ou quanto muito se determinada palavra é um palavrão. Isto porque a tarefa em causa é *filtrar* mensagens — daí o grande foco no léxico. O nosso objetivo é *reconhecer* os palavrões, isto é, se este aglomerado de símbolos for um palavrão, que palavrão é?

A diferença que este passo pode fazer no âmbito da análise automática de comunidades, sentimentos, opiniões ou outra forma de estudo assente sobre conteúdos textuais gerados por utilizadores, é que se conserva (recupera, até se poderá dizer) informação adicional emitida pelo autor. Se alguém se refere ou dirige a uma pessoa usando um palavrão — vamos imaginar, a título de tecer um cenário, que estudamos a imagem de uma figura pública nas redes sociais, e sabemos apenas que alguém usou um palavrão em referência a essa figura pública — então há muito pouca informação que se pode inferir. No entanto, se reconhecermos o palavrão podemos perceber que pode estar a referenciá-la como homossexual, como traída na sua relação amorosa, como desprovida de valor ou como sendo sexualmente desejada, para referir apenas alguns exemplos. O que procuramos fazer é aumentar a cobertura de situações desse género com que conseguimos lidar. Observando os números que dispomos sobre a ofuscação de mensagens na Internet (Laboreiro e Oliveira, 2014), ilustrados na Figura 5, estes números podem, em alguns casos, ser significativos.

Existem outras tarefas semelhantes à desofuscação, na medida em que visam "regularizar" textos, como é o caso da correção de erros ortográficos, que trata mudanças de grafia acidentais sobre um léxico muito mais vasto. A desofuscação e a correção de erros são, por natureza, tarefas de *normalização*, ou seja, tornar a grafia das palavras previsível, constante e com o mínimo de variação possível.

2 Estado da arte

Os palavrões, do ponto de vista da informática, têm sido abordado poucas vezes e com poucos avanços. Numa análise dos principais problemas dos sistemas de deteção de palavrões (Sood, Antin e Churchill, 2012a), os autores avançam que a aparente facilidade desta tarefa, que é muito difícil em contextos reais, pode ter contribuído para que esta área tenha sido negligenciada em termos de investigação.

Iremos abordar três vertentes associadas a este problema, que cobrem essencialmente os trabalhos científicos que tratam a questão dos palavrões: trabalhos simples baseada em listas, o tratamento de mensagens insultuosas ou ofensivas, e estudos linguísticos que abordam o uso de palavrões. No entanto nenhum destes trabalhos se propôs a tratar o mesmo problema que abordamos neste trabalho.

2.0.1 Trabalhos baseados em listas

Estes são os trabalhos mais simples. O problema que se tentava resolver pode ser descrito de forma muito sintética: há uma lista de palavras que deviam *nunca* surgir em mensagens lidas ou escritas pelos utilizadores. O objetivo era pois filtrar (ou no mínimo identificar) estas mensagens corretamente. Na verdade, a responsabilidade residia toda na componente lexical, e como tal o funcionamento destes sistemas era muito elementar.

Inicialmente estas aplicações destinavam-se a proteger as crianças da exposição a linguagem rude na Internet (Jacob et al., 1999), e a manter a devida compostura nas empresas (Cohen, 1998) — os anos 90 trouxeram novas ferramentas eletrónicas que, aparentemente, inspiravam informalidades indevidas nos locais de trabalho.

A patente norte-americana identificada pelo número 5 796 948 (Cohen, 1998) documenta a sua invenção como "um método para a interseção de comunicações ou correspondência em rede, que contenha palavras ofensivas ou profanas, fragmentos de palavras, frases, parágrafos ou outra unidade de linguagem que possa ser formulada em qualquer linguagem natural (...) ou artificial." O documento é bastante omisso na descrição do método usado para reconhecer esse tipo de ofensas (não era este o foco principal da patente), mas tudo indica que consiste unicamente em encontrar substrings na mensagem que estejam enunciadas numa lista pré-determinada.

No ano seguinte, Jacob et al. abordaram um sistema de classificação automática de websites (Jacob et al., 1999), que passava por classificá-los, referindo a sua adequação à exploração por crianças e jovens em várias faixas etárias. A avaliação dos websites era feita por humanos, o que os autores admitem poder vir a criar a um problema de escalabilidade. No entanto o artigo mencionava que as abordagens baseadas em listas de palavras-chave têm uma aplicação demasiado "cega" das regras, só funcionam com texto, e mesmo o texto pode não ser bem validado. Com isto queremos salientar que já na altura se procurava uma alternativa para estes métodos, mas passaram já 15 anos sem que melhores soluções se tenham implantado.

Com a massificação do acesso à Internet, que aconteceu mais tarde, olhou-se de novo e com mais atenção para a questão de proteger os utilizadores do conteúdos indesejados. O foco expandiu-se, e passou de palavrões para todo o tipo de insultos escritos.

2.0.2 Lidar com mensagens insultuosas

A procura de mensagens insultuosas tem alguma afinidade com a identificação de palavrões. Aliás, todos estes sistemas integram um léxico de palavrões a que recorrem durante o seu funcionamento, e mostram uma aplicação direta do reconhecimento deste vocabulário.

Já em 1997 Spertus descrevia *Smokey*, um sistema que detetava "mensagens hostis" (Spertus, 1997), também chamadas de "flames" em inglês. Como refere o autor, estas mensagens não são o mesmo que "expressões obscenas", já que apenas 12% continham vulgaridades, e mais de um terço das mensagens contendo vulgaridade não eram consideradas abusivas — os palavrões podem ser usados como expletivo, por exemplo.

O Smokey fazia uso de uma série de regras fixas e pré-definidas, algumas das quais recorrendo a uma lista de palavrões. Por exemplo, as regras que lidam com palavrões são ativadas quando um dos palavrões considerados é encontrado, mas opera de forma diferente caso um "vilão" seja mencionado na mesma frase (um "vilão" é uma entidade — como um político ou personagem de uma série de TV — da qual é usual dizer mal na comunidade em questão, e como tal o uso de palavrões é *esperado*).

Spertus parece ter evitado o texto ruidoso no seu trabalho, visto referir que removeu "mensagens sem sentido (alguém pressionando teclas ao calhas) das coleções", e posteriormente, quando discute as limitações do sistema, refere uma mensagem que "não foi detetada devido à sua tipografia pouco usual", que era "G E T O V E R I T" (em português seria algo como "S U P E R A I S S O").

Mais tarde, em 2008, Mahmud, Ahmed e Khan apresentaram a sua abordagem à deteção automática de mensagens hostis e insultos recorrendo a informação semântica (Mahmud, Ahmed e Khan, 2008). Os autores admitem as limitações do seu sistema, afirmando que "não lidava ainda com texto erróneo, tal como a má colocação da vírgula, sinais de pontuação não correspondidos [à falta de esclarecimento dos autores, assumimos que se referiam a símbolos usados normalmente aos pares, como aspas ou parêntesis, dos quais só se encontrou um no texto], etc.".

Em 2010, Razavi et al. desenvolveram um sistema de classificação multi-nível para a deteção de textos abusivos (Razavi et al., 2010), mas no estudo os autores descartaram das mensagens todos os símbolos que não fossem alfabéticos ou de "pontuação expressiva".

Também nesse ano, Xu e Zhu propuseram uma abordagem para filtragem de mensagens ofensivas que operava ao nível das frases (Xu e Zhu, 2010). O seu objetivo era remover o conteúdo desagradável mantendo a integridade global da frase. O leitor, idealmente, não daria pela operação sobre a mensagem original. Referem que a ofuscação traz um problema difícil que deve ser tratado como uma questão específica.

Dois anos depois, Xiang et al. procuram detetar *tweets* ofensivos recorrendo a uma abordagem de *bootstrapping* (Xiang et al., 2012). Consideram apenas palavras compostas por letras e os símbolos - e '.

A título de resumo, enquanto que os trabalhos da subseção anterior descreviam trabalhos que consistiam em verificar se uma palavra constava numa lista de palavrões (Cohen, 1998; Jacob et al., 1999), aqui falámos de vários trabalhos que usam esse mecanismo para procurar ou filtrar insultos. Foram diversos os métodos adotados para este fim: regras pré-definidas (Spertus, 1997), análise semântica (Mahmud, Ahmed e Khan, 2008; Xu e Zhu, 2010) ou classificação automática (Razavi et al., 2010; Xiang et al., 2012). No entanto, tenham sido usados métodos mais simples ou mais sofisticados, nenhum destes trabalhos procuram lidar com palavras que não estejam bem escritas, sejam palavrões ou não. Isto compreende-se, já que o foco do trabalho situava-se mais ao nível do reconhecimento semântico das mensagens. De seguida iremos abordar estudos que tratam o tema dos palavrões em grandes coleções de mensagens, e têm em atenção a questão da grafia variável.

2.0.3 Estudos em coleções de mensagens

Alguns trabalhos focaram-se mais no estudo do uso ou no reconhecimento de palavrões. No entanto, a fim de podermos comparar os seus resultados (e saber o que podemos esperar da aplicação dos seus métodos) necessitamos de considerar algumas decisões que foram tomadas no âmbito do trabalho que foi desenvolvido. Nomeadamente é relevante especificar como o léxico foi obtido e como foi feito o reconhecimento dos palavrões em cada um dos casos estudados.

MySpace Na análise da sua coleção de *homepages* do MySpace (Thelwall, 2008), foram compiladas duas listas de palavrões. Para a lista britânica, Thelwall começou com palavrões que constavam no guia oficial da BBC, acrescentando variantes comuns, como sufixos. De seguida encontrou palavras formadas por aglutinação, procurando nos textos pelo tronco das palavras já consideradas. Devido ao grande número de resultados obtidos desta forma, excluiu as que surgiam em menos de 0,1% dos perfis. De seguida procurou no léxico por variantes gráficas das palavras (acidentais ou propositadas). Por fim, acrescentou um pequeno número de palavrões bem conhecidos.

Palavrões...
Muito fortes: *cunt, motherfuckin, motherfucking, muthafucker, muthafuckin, mutherfucker*
Fortes: *fuck, fucked, fucken, fucker, fuckin, fucking, fuckstick*
Médios: *aresehole, asshole, bastard, follock, piss, pissin, pissing, shagged, shagging, twat, wank, wanker, wanking*

Tabela 4: Palavrões considerados por Mike Thelwall na sua análise principal dos perfis do MySpace.

Para a lista criada só para os utilizadores dos EUA o autor baseou-se nas "sete palavras sujas" ("seven dirty words"), uma lista de palavras conhecidas como estando proibidas de serem transmitidas em sinal aberto pela Comissão Federal de Comunicações desse país[2], sendo elas *shit, piss, fuck, cunt, cocksucker, motherfucker* e *tits*.

O autor não usou todas as palavras nas experiências, fazendo uso de apenas 6 palavrões muito fortes, 7 palavrões fortes e 13 palavrões de intensidade média. Não é perfeitamente claro no artigo quais os palavrões que co-existiam nas duas listas, listando a Tabela 4 as palavras numa forma agregada.

Yahoo! Buzz Sood, Antin e Churchill reutilizaram a coleção que tinham anotado para a deteção de insultos pessoais (Sood, Churchill e Antin, 2011) para a deteção de palavrões. Primeiro abordaram os problemas associados ao uso de sistemas de correspondência direta com listas de palavrões (Sood, Antin e Churchill, 2012a), onde afirmam que as listas são fáceis de contornar (recorrendo à ofuscação), são de adaptação difícil (não conseguem lidar com abreviaturas e erros), e que dificilmente conseguem acomodar mais que um domínio.

Recorrendo à suas 6354 mensagens anotadas e a duas listas de palavrões disponíveis on-line, os autores demonstram a inadequação da corres-

[2]Apesar desta lista ser bem conhecida, é discutido se alguma vez foi criada uma enunciação formal de palavras inapropriadas para transmissão em televisão.

pondência direta em tarefas de identificação de palavrões. Os autores procuraram identificar de forma automática quais as mensagens que continham palavrões, tendo por base duas listas de palavrões disponíveis on-line. O melhor resultado que obtiveram com este método (usando a medida F1 como principal critério de avaliação) foi precisão 0,53, cobertura 0,40 e F1 0,46, recorrendo à radicalização (*stemming*).

Avaliando as palavras que aparentavam ser as mais discriminatórias na distinção entre uma mensagem ter ou não ter palavrões (recordamos que a granularidade desta anotação é ao nível da mensagem), apenas 8 em 33 palavras estavam bem escritas e metade das palavras na lista usavam símbolos não alfabéticos para fins de ofuscação. Para ilustrar como o uso destes métodos é difícil de resolver de forma automática, 40% dos usos do símbolo "@" na sua coleção destinavam-se a ofuscar palavrões, enquanto os restantes 60% se distribuíam entre endereços de email, direcionar mensagens a utilizadores, e outros fins não insultuosos.

Um trabalho subsequente dos mesmos autores (Sood, Antin e Churchill, 2012b) propõe duas soluções para este problema. A primeira recorre à distância de Levenshtein (Levenshtein, 1966) para comparar as palavras no texto com os palavrões na lista. Esta comparação é feita apenas no caso da palavra não ser reconhecida diretamente como palavrão ou como uma palavra no dicionário de inglês. A palavra é considerada um palavrão se o número de edições (o número total de operações de inserção, remoção e/ou alteração) igualar o número de "sinais de pontuação[3]" no termo ou for inferior a um certo valor que varia com o seu comprimento.

A segunda solução apresentada faz uso de SVMs (Support Vector Machines), com uma abordagem de "bag of words" baseada em bigramas, com vetores binários representando a presença ou ausência destes bigramas, e empregando um *kernel* linear. Este classificador, após ter sido treinado com 1/10 das mensagens numa sequência de testes de validação cruzada, prevê se cada uma das restantes mensagens contém ou não algum palavrão.

Individualmente as SVMs obtiveram precisão 0,84, cobertura 0,42 e F1 0,56, enquanto que a combinação (por disjunção) das SVMs, Levenshtein e uma das listas empregues no trabalho anterior resultou em precisão 0,62, cobertura 0,64 e F1 0,63. Ou seja, reconheceu mais palavrões, mas à custa dos falsos positivos. Estes valores são, relembra-se, relativos à classificações binária da presença de palavrões em *mensagens*, e não à identificação de palavras que são palavrões.

Twitter Wang et al. elaboraram um léxico mais elaborado que o de Sood, Antin e Churchill na sua análise da plataforma de *microblogging* (Wang et al., 2014). Começaram por juntar várias listas de palavrões existente na Internet, que foram compiladas com o propósito de servirem sistemas de filtragem de palavrões. Retiveram apenas as palavras em inglês que são usadas principalmente de forma ofensiva. Esta filtragem foi feita manualmente, resultando em 788 palavras, contendo também variações gráficas usadas com o intuito de ofuscar as palavras.

Os autores procuraram inovar também no reconhecimento de palavrões ofuscados. Para todas as palavras que não são constam no léxico de palavrões, procederam primeiro à normalização das palavras removendo letras repetidas. Depois substituíram algarismos e símbolos pelas letras mais semelhantes graficamente. Por fim fizeram uso da distância de Levenshtein para efetuar a correspondência com a lista de palavrões. Consideraram que havia correspondência se a distância de edição for igual ao número de símbolos de máscara (símbolos usados para disfarçar palavras, normalmente em substituição das letras), mais concretamente, os seguintes sete símbolos: _ % - . # \ ' . Fica por esclarecer se, por exemplo, "####" seria interpretável como um palavrão, ou porque os asteriscos não foram considerados (assim como, possivelmente, no trabalho de Sood, Antin e Churchill, caso tenham mesmo considerado apenas sinais de pontuação).

Na sua experiência, os autores anotaram manualmente uma amostra de 1000 tweets escolhidos aleatoriamente da sua coleção, como contendo ou não palavrões. Usando o método descrito acima obtiveram uma precisão de 0,99, cobertura de 0,72 e medida F1 de 0,83 na classificação automática das mensagens. Ou seja, muito poucas mensagens sem palavrões foram reconhecidas erradamente como tendo um, e a grande maioria das mensagens que efetivamente tinham palavrões foi reconhecida como tal.

SAPO Desporto A análise de 2500 mensagens extraídas do SAPO Desporto, um website português de notícias desportivas, que foram anotadas manualmente (Laboreiro e Oliveira, 2014), revelou que uma em cada cinco mensagens continha um ou mais palavrões.

Apesar do tamanho ser relativamente reduzido, há alguns aspetos a salientar deste traba-

[3]É possível que os autores se estejam a referir a símbolos não alfanuméricos e não espaços em branco.

lho que o tornam, de uma forma ou de outra, relevante.

Primeiro, todas as mensagens encontram-se anotadas com granularidade mais fina que nos corpus anotados que mencionámos até aqui, ou seja, é anotado ao nível da palavra invés de o ser ao nível da mensagem, que é o mais comum. Consequentemente é possível calcular a medida de cobertura assim como trabalhar com um léxico perfeito — isto é, todos os falsos negativos (palavrões que não são encontrados) correspondem a falhas no processo de reconhecimento e nunca a um palavrão que, por infeliz acaso, não foi considerado durante elaboração do léxico. Na Secção 1.3 vimos como é fácil isso acontecer.

Em segundo lugar, esta coleção está disponível para uso livre em licença aberta[4]. Isto possibilita que várias abordagens e algoritmos sejam comparados e medidos contra o mesmo padrão, o que é essencial na ciência. Tanto quanto é do conhecimento dos autores, este é o único *corpus* dedicado ao estudo dos palavrões disponível desta forma.

Finalmente é importante consagrar a língua portuguesa que, apesar de ser uma das línguas mais populares na Internet, tem sido menosprezada neste tema. Todas as mensagens estão escritas em português.

Todos estes motivos levaram a que esta coleção fosse escolhida como base para o nosso estudo, que passamos a descrever, começando até por apresentar alguns aspetos relevantes do corpus SAPO Desporto.

3 A coleção de mensagens utilizada

O objetivo deste trabalho consiste na identificação e reconhecimento de palavrões que foram escritos de uma forma que dificulta a sua compreensão. Deste ponto em diante será apresentado o trabalho que foi desenvolvido nesse sentido. Mais concretamente, será avaliada a adequabilidade dos métodos geralmente utilizados para a identificação de palavrões (ou seja, dizer se é palavrão ou não), e a sua capacidade no reconhecimento dos mesmos (dizer qual foi o palavrão encontrado).

Iremos aproveitar esta secção para apresentar os dados que foram utilizados na avaliação, já que tal permitirá ao leitor perceber melhor o problema da ofuscação e as dificuldades da sua remoção.

Como foi dito anteriormente, optou-se por usar uma versão revista da coleção SAPO Desporto, uma coleção destinada ao reconhecimento de palavrões. A revisão dos dados permitiu resolver algumas inconsistências pontuais na anotação. Dado que a coleção SAPO Desporto encontra-se já devidamente descrita (Laboreiro e Oliveira, 2014), apenas alguns pontos mais relevantes serão aqui destacados.

A fonte original destas mensagens foi o website de notícias SAPO Desporto, que foi escolhida exatamente devido à profusão de palavrões existente nos comentários. É interessante observar esta oposição (por certo cultural) com o Yahoo! Buzz, onde o desporto foi a categoria com menos palavrões observados (Figura 3).

Nas 2500 mensagens foram assinalados 776 usos de palavrões no total (palavrões repetidos na mesma mensagem só contam uma vez se forem escritos da mesma forma). E apesar dos palavrões estarem presentes em mais de 20% das mensagens, quase 30% das mensagens com palavrões têm mais do que um. A forma como estes se acumulam está ilustrada na Figura 4.

Como o SAPO empregava um sistema (muito simples) de filtragem de palavrões, contorná-lo tornou-se uma prática comum. Na verdade os 5 palavrões mais frequentes só se encontram em forma ofuscada, e o filtro do SAPO bloqueava 23 dos 30 palavrões mais vistos (ou seja, continuam a ser muito utilizados apesar da filtragem). Esta filtragem proporcionou numa amostra rica em variedade de palavrões, através da procura de palavras ausentes do léxico do SAPO. Foram identificados 40 palavrões-base, expandidos para 111 quando considerando variações (indicados na Tabela 7, em apêndice). Mas o registo apresenta-se rico também em métodos de ofuscação via grafias alternativas, como ilustrado na Figura 5. Encontrou-se uma média de 3 grafias para cada um dos 111 palavrões. Estes métodos de ofuscação foram compilados em 17 categorias por Laboreiro e Oliveira, que destacaram as seguintes como sendo as mais comuns:

- A substituição de uma letra por outra letra (normalmente mantendo a pronúncia da palavra inalterada), por exemplo: “ku”;
- A repetição de letras, como em “puuta”;
- O uso de sinais de pontuação ou espaços como separadores de letras, e.g. “m.e.r.d.a”;
- O recurso a algarismos para tomar o lugar de letras graficamente semelhantes, como visto em “m3rda”;
- A substituição de letras por símbolos, graficamente semelhantes ou não, como nas palavras “put@” ou “p*ta”; e

[4] `http://labs.sapo.pt/2014/05/obfuscation-dataset/` visto em 2014-12-20

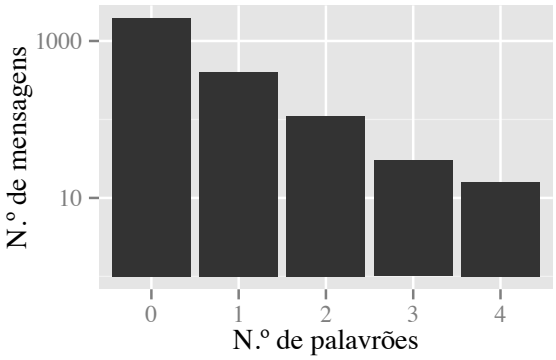

Figura 4: Número de mensagens contendo determinado número de palavrões assinalados. Grafias repetidas na mesma mensagem não foram contabilizadas.

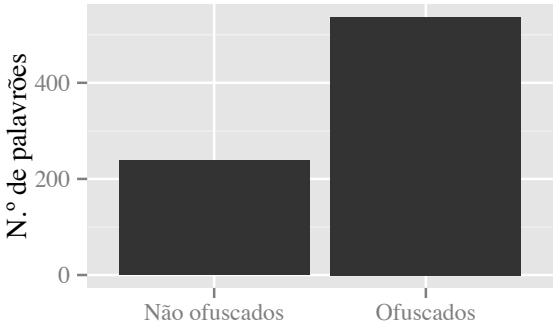

Figura 5: Comparação do número palavrões não ofuscados com o número de palavrões ofuscados. Grafias repetidas na mesma mensagem não foram contabilizadas.

- A supressão ou modificação de acentos, como em "cabrao".

Ao ofuscarem palavrões os utilizadores parecem preferir a substituição de caracteres mais do que a inserção ou remoção de símbolos, possivelmente porque a substituição preserva a dimensão da palavra, o que funciona como uma pista para a sua descodificação.

Da anotação da coleção resultam três derivados importantes. Em primeiro lugar, estão assinaladas quais as palavras no texto que são palavrões, o que é necessário para a tarefa de identificação dos palavrões. Resulta também a correspondência entre as palavras ofuscadas e os palavrões que representam, que é informação relevante para a tarefa de reconhecimento. É também daqui que se recolhe o léxico, que é simplesmente a lista de todos os palavrões encontrados e ao mesmo tempo todos os palavrões que nos interessam. Por fim, desta correspondência é possível extrair informação sobre *como* é feito o processo de ofuscação. Esta informação foi usada apenas para o estudo geral da coleção (Laboreiro e Oliveira, 2014), mas a sua importância não se esgota aí. Na secção 7 são apresentadas algumas propostas como trabalho futuro para o seu melhor aproveitamento.

De seguida tratamos dos métodos empregues para a desofuscação automática de palavrões.

4 Metodologia

Descrevemos já algumas estratégias empregues na identificação de palavrões, como é o caso da procura das palavras vistas num léxico ou o cálculo da distância de Levenshtein. Para avaliar as abordagens mais comuns e verificar se conseguem identificar os palavrões que não estão escritos na sua forma canónica, procedemos à divisão da tarefa em duas componentes: pré-processamento, onde manipulamos o texto original de forma a reduzir o seu ruído e o reconhecimento propriamente dito do palavrão, que se espera que fique mais facilitado devido ao tratamento prévio do texto.

A tarefa de pré-processamento é composta pela atomização e normalização, descritas de seguida, enquanto os métodos usados para fazer a correspondência entre as palavras vistas e os palavrões no léxico são descritos mais à frente, na Secção 4.3.

4.1 A atomização

A atomização (em inglês, *tokenization*) é o processo de dividir o texto em átomos lógicos para processamento individual, tradicionalmente palavras, pontuação e números. Hoje em dia há mais tipos de átomos que necessitam de ser tomados em consideração, como URLs, endereços de email e smileys. A atomização deficiente é suscetível de causar a perda de informação em situações em que o texto é ruidoso (Laboreiro et al., 2010).

Dois métodos de atomização foram usados no nosso trabalho. Um simples, baseado em poucas regras, que delimita os átomos na fronteira entre carateres alfanuméricos e carateres não alfanuméricos (como espaço em branco ou pontuação). Mais concretamente fez-se uso da expressão regular `\b[\w\d_-]+\b` para identificar os átomos relevantes.

Os delimitadores usados por esta expressão regular são semelhantes aos empregues nos atomizadores mais antigos, como é o caso do Penn Tre-

ebank Tokenizer,[5] e apesar de ser fácil expandir esta expressão para abarcar situações mais complexas, o objetivo era mesmo a simplicidade.

O outro atomizador utilizado foi desenvolvido tendo em consideração texto mais ruidoso, como o do Twitter, e está disponível on-line para uso livre (Laboreiro et al., 2010), na coleção "Sylvester"[6]. Este atomizador usa uma SVM treinada com mensagens do Twitter anotadas manualmente e visa encontrar os pontos de descontinuidade (entre carateres alfanuméricos e não alfanuméricos) ideais para separar o texto. Espera-se que esta ferramenta consiga lidar melhor com palavras ofuscadas, como "m.rda", sem que sejam separadas em três átomos, dada a abordagem mais sofisticada e os testes com textos ruidosos.

Dado que as etapas de processamento subsequentes lidam com os átomos a título individual, serão necessárias condições muito favoráveis para recuperar um palavrão cortado durante o processo de atomização. No entanto, algum símbolo extra que permaneça agregado à palavra pode comprometer a correspondência, ao diminuir a similaridade com o palavrão. Este não reconhecimento é mais provável se múltiplas palavras estiverem "agarradas", como por exemplo, "se-o-texto-for-escrito-desta-forma". O equilíbrio entre a decisão de cortar ou manter um símbolo ao formar átomos é algo ténue, e por esse motivo testamos dois métodos.

Sabendo que a nossa tarefa de atomização se foca unicamente em *dividir* o texto em átomos, e nunca trata de *agregar* sequências de letras separadas pelo caráter espaço (que é o separador de átomos *de facto* nos nossos textos), pode considerar-se à partida que as palavras ofuscadas com recurso a este símbolo estão irremediavelmente mal atomizadas. Olhando para os nossos dados encontrámos 37 instâncias em que isso acontece, muitas das vezes usando mais do que um espaço para esconder a palavra. Será muito difícil reconhecer algum destes palavrões fazendo uso dos métodos descritos neste trabalho.

Uma situação também problemática (embora menos) é o uso de pontuação no processo de ofuscação, visto que é um excelente candidato a definir fronteiras de átomos. Na nossa coleção encontrámos 48 instâncias em que isso acontece.

É importante realçar que os objetivos deste trabalho passam por fazer uma correspondência perfeita entre o palavrão ofuscado e o palavrão desofuscado, o que exige o reconhecimento do átomo correto na mensagem. Ou seja, pretende-se que ao fazer a operação de clarificação da mensagem não se deixe "ruído" adicional na mesma, nem, por outro lado, se suprimam carateres necessários. Por exemplo, imagine-se que a mensagem contém o palavrão "m.rda" e o atomizador age de forma incorreta, produzindo "m . rda". Se o reconhecedor conseguir fazer a correspondência entre "rda" e "merda", consideramos como estando errado, visto que, ao corrigir a mensagem fica o texto "m . rda". Portanto, encontrar o palavrão correto não é suficiente.

4.2 A normalização

A normalização procura reduzir o número de formas diferentes como se escrevem as palavras. Ao restringir o alfabeto com que se trabalha, reduz-se também a complexidade do mapeamento entre o texto e o léxico. Por exemplo, um erro encontrado frequentemente consiste na omissão do acento em algumas palavras, duplicando assim o número de grafias a considerar para reconhecer essas palavras. Se abolirmos os acentos, ambas as representações (com e sem acento) convergem numa só forma (sem acento). Este processo é de certa forma análogo a converter *strings* para minúsculas antes de as comparar.

É certo que este processo também pode originar efeitos colaterais, promovendo a confusão entre palavras diferentes. Contudo, o número de casos que devem afetar o nosso léxico será insignificante no máximo (é muito improvável que alguém invoque *cágados* num website dedicado ao desporto), e como tal os benefícios deverão compensar largamente os problemas que possam surgir. As experiências deverão verificar se isto é verdade.

Definiram-se quatro níveis de normalização:

Nenhum em que nada é alterado, servindo apenas para efeitos de comparação;

Mínimo onde se efetuam modificações simples:

- As letras são todas convertidas para minúsculas;
- Os acentos e cedilhas são removidos;
- Os espaços em branco repetidos são condensados;
- Referências a utilizadores, como "@xpto", são trocadas por um símbolo;
- URLs são substituídos por um símbolo;
- As hashtags perdem o caráter "#" e passam a ser palavras normais;

[5] `http://www.cis.upenn.edu/~treebank/tokenizer.sed` visto em 2014-12-20

[6] `http://labs.sapo.pt/2011/11/sylvester-ugc-tokenizer/` visto em 2014-12-20

Símbolos	Letra	Símbolos	Letra
0	o	5	s
1	i	6	g
2	z	7	t
3 €	e	8	b
4 @	a	9	g

Tabela 5: Tabela de substituição de símbolos por letras através de equivalência gráfica.

Básico onde se efetuam as tarefas do nível "Mínimo", mais:

- Os números são transformados nas letras graficamente mais semelhantes, como enunciado na Tabela 5;

Máximo onde se efetuam as tarefas do nível mínimo, acrescido do seguinte:

- Vários símbolos mais incomuns são convertidos para os símbolos da tabela ASCII, usando como base uma tabela de carateres confundíveis, obtida do Consórcio Unicode[7].

Estas transformações são aplicadas também ao conteúdo do léxico, a fim de possibilitar a devida correspondência na fase do reconhecimento.

4.3 O reconhecimento

O processo de reconhecimento é responsável por fazer corresponder os átomos encontrados no texto (como palavras) com os elementos constantes no léxico de palavrões. Usámos três métodos de correspondência amplamente conhecidos:

Igual que indica apenas se ambas as sequências de carateres são idênticas;

Substring que indica se a palavra no léxico está contida no átomo avaliado; e

Levenshtein que mede o número de operações de edição do algoritmo de Levenshtein necessárias para transformar o átomo em questão no palavrão com que está a ser comparado (Levenshtein, 1966).

Os dois últimos métodos aceitam um parâmetro, compreendido entre 0 e 1, que define o grau mínimo de semelhança que é exigido para reconhecer a correspondência. O valor 0 é o mais permissivo enquanto o 1 é mais exigente.

No caso do método "Substring", o parâmetro define a proporção da palavra que deve ser idêntica entre os dois valores comparados, a fim de idealmente permitir encontrar "cu" em "cus" mas não em "curvilíneo", por exemplo. No caso do método "Levenshtein" o parâmetro corresponde ao nível de semelhança, mas de acordo com a fórmula $1 - E/C$, em que E é o número de edições e C é o comprimento do átomo analisado. Em ambos os métodos o comportamento converge para o do "Igual" à medida que o valor do parâmetro se aproxima de 1.

[7] `http://www.unicode.org/review/pri273/` visto em 2014-12-20

5 Descrição da experiência

Iremos agora elaborar sobre a estrutura do sistema de avaliação que foi concebido para medir o contributo dos diversos métodos atrás descritos na tarefa de desofuscação de palavrões. Esta experiência prevê a avaliação de todas as combinações de atomizadores, normalizadores e reconhecedores. Desta forma é possível identificar aqueles que são mais ou menos adequados, tal como calcular as diferenças de desempenho que cada um potencia.

Usou-se a coleção de 2500 mensagens anotadas do SAPO Desporto que possui todos os palavrões já assinalados. No entanto os testes não foram adaptados às condições particulares deste corpus, ou seja, não se tomou partido do conhecimento prévio dos métodos de ofuscação mais comuns.

Para cada método de reconhecimento configurável testaram-se os valores de semelhança em intervalos de 0,1, entre 0,1 e 1,0, que corresponde ao nível máximo de exigência. O algoritmo de Levenshtein foi configurado para atribuir o mesmo peso às suas três operações de edição.

O desempenho foi quantificado recorrendo às medidas de precisão, cobertura e a média harmónica de ambas, conhecida como F1. Considera-se um sucesso apenas um mapeamento correto entre o átomo que é observado no texto e o palavrão que ele representa.

Deve ser salientado desde já que o teste é bastante exigente, pois esta avaliação exige que num caso como "seus grandes cabr.." consiga resolver o palavrão correto assim como o seu género e número. Caso contrário é declarado como um erro (falso positivo), visto que a identificação de palavrões não é contabilizada só por si.

Parte da dificuldade inerente à tarefa de desofuscação prende-se com os casos em que pode existir mais do que uma solução possível. Por exemplo, "m3rda" pode ser desofuscado como "merda" ou "morda", e "p*ta" pode resultar em "pata" ou "puta". Para um computador é efetivamente igual, já que se trata apenas de trocar um símbolo por outro.

Uma possível solução possível para este problema seria optar pela solução mais frequente nos textos, dado que tanto "morda" como "pata" são palavras algo improváveis. Acontece que "merda" e "puta" são inexistentes nos textos da coleção, visto que são palavras detetadas pelo sistema de filtros do SAPO. Optou-se por assumir que no presente contexto não há motivo para os utilizadores ofuscarem palavras que não sejam ofensivas, e como tal, as soluções que são palavrões são mais valorizadas.

6 Resultados e análise

Apresentar todos os resultados que foram obtidos no decorrer da nossa experiência seria muito confuso e ineficiente. Por isso optámos por focar a nossa atenção, primeiro no tema do pré-processamento e depois no processamento em si, convergindo na melhor solução que foi alcançada.

Assim sendo, comecemos por observar como as medidas de precisão e de cobertura são afetados pela atomização e normalização, como representado na Figura 6, que engloba seis gráficos. A precisão é representada nos eixos das abcissas enquanto que a cobertura ocupa o eixo das ordenadas. Cada coluna de gráficos apresenta os resultados fixando cada um dos métodos de atomização, enquanto que as linhas representam cada um dos métodos de normalização. Cada método de reconhecimento surge representado por um símbolo diferente, como indicado na legenda, e apresentam os resultados referentes à variação dos seus parâmetros. O método igual surge menos vezes no gráfico por dois motivos: primeiro, não tem parâmetro, e segundo, porque é menos sensível ao pré-processamento, e como tal os seus pontos tendem a estar sobrepostos.

6.1 A Atomização

Começando pela atomização, podemos notar como a escolha do método utilizado não tem muita influência nos valores obtidos, visto que o método "Sylvester" parece trazer nenhum ganho sobre o "Simples". Na verdade, se observado com atenção, nota-se que os resultados são até ligeiramente piores, quer em precisão quer em cobertura.

Dos 776 palavrões assinalados, 134 falham devido à atomização mal feita com o método "Simples" e 148 com o método "Sylvester". Estes casos correspondem a falsos negativos provocados quer por situações complexas que a atomização não resolve (e.g. ofuscação usando espaços), quer por problemas que a atomização

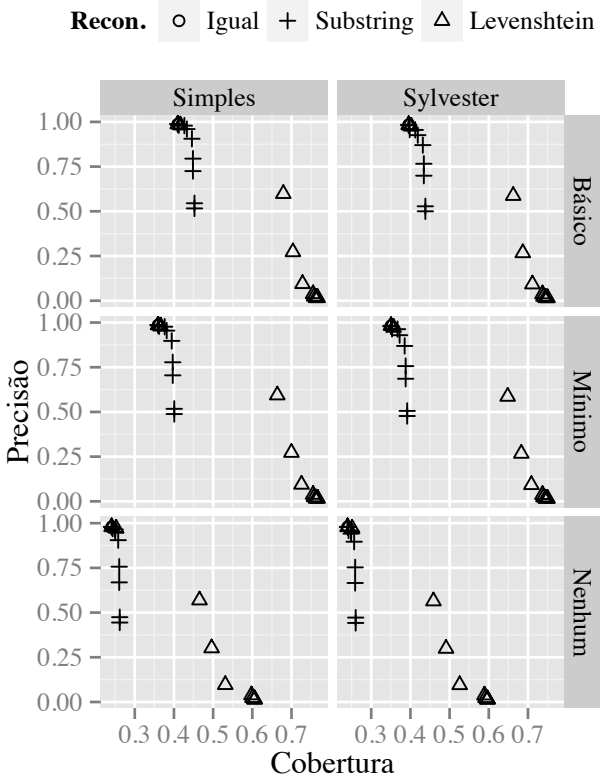

Figura 6: Comparação do desempenho dos dois atomizadores (as colunas) e três métodos de normalização do texto (as linhas) com todos os métodos de reconhecimento.

cria (e.g. ofuscação usando pontuação). De qualquer forma, nestas situações, que correspondem a 17–19% de cobertura, o átomo certo nunca chegará a ser devidamente avaliado, e estão desde já perdidas.

6.2 A Normalização

Por seu lado a normalização consegue fazer muito pelo reconhecimento de vários palavrões que passariam despercebidos de outra forma (os falsos negativos), e cada nível de operação que se adiciona parece contribuir neste sentido.

Ainda assim, a normalização mais completa (a que chamámos o nível "Máximo") não difere do nível "Mínimo", devido à ausência de substituições mais elaboradas nas mensagens usadas. Por esse motivo não surge ilustrada na Figura 6. Possivelmente o filtro do SAPO, que era bastante simples, permitia aos utilizadores o mesmo resultado trocando um "e" por "3" ou por "€", que estão facilmente acessíveis no teclado, o que não acontece com o "ε". Por esse motivo não destacamos os resultados do nível de normalização "Máximo" durante a nossa análise.

Os principais ganhos de cobertura foram obtidos no nível "Mínimo" de normalização, com

benefícios mais modestos fornecidos pelo nível "Básico". O método de reconhecimento "Levenshtein" foi o reconhecedor que beneficiou mais deste pré-processamento.

Focando-nos agora na precisão, não foram notadas diferenças significativas entre os níveis de normalização "Mínimo" e "Básico". O método de reconhecimento "Igual" permaneceu quase inafetado pela mudança de normalização. Já com o "Substring" nota-se que os parâmetros mais permissivos melhoram com a normalização, permitindo-lhe ascender acima dos 0.50 quando esta é mais ativa. Por fim, o algoritmo de Levenshtein apresenta ganhos muito tímidos de precisão, possivelmente devido à sua capacidade de tolerância superior.

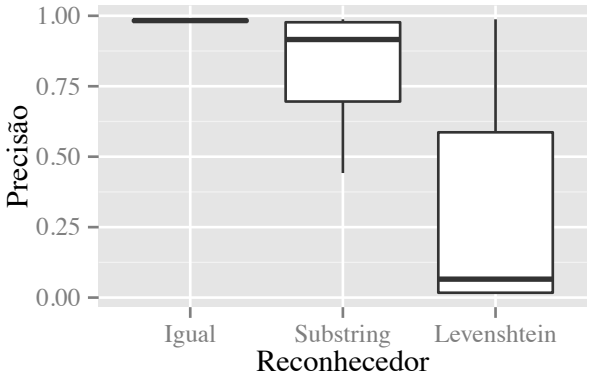

Figura 7: Comparação do desempenho dos três reconhecedores em termos de precisão.

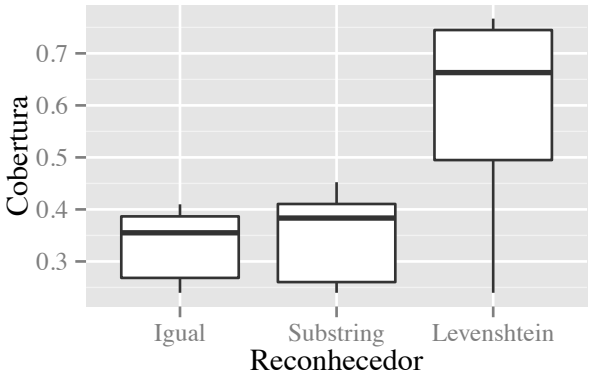

Figura 8: Comparação do desempenho dos três reconhecedores em termos de cobertura.

6.3 O Reconhecimento

Continuamos a nossa análise focando-nos apenas na atomização com o método "Simples" e normalização "Básica", cuja combinação foi a que proporcionou os melhores resultados. Iremos primeiro comparar o desempenho dos três reconhecedores, e de seguida analisar como se comportam perante os diversos valores de tolerância.

6.3.1 Os Três Reconhecedores

A Figura 7 mostra as medidas de precisão obtidas por todos os reconhecedores, enquanto a Figura 8 representa os valores de cobertura alcançados com os vários valores de parâmetro. Os gráficos de caixa e bigodes representam o primeiro, segundo e terceiro quartis com linhas horizontais, sendo a mediana destacada. As linhas verticais representam os valores máximos e mínimos, não havendo qualquer valor além de 1,5 vezes a distância inter-quartil.

Podemos confirmar como os métodos "Igual" e "Substring" são precisos, pouco dados a falsos positivos, mas rígidos e incapazes de lidar com ofuscações mais elaboradas. O método "Igual" é bastante fiável no que toca à precisão, mas a sua cobertura é bastante baixa. É curioso também comparar o potencial ganho de cobertura que uma comparação por substring pode trazer sobre uma igualdade, com o potencial risco que isso pode trazer ao nível da precisão.

Por seu lado, o método de Levenshtein permite encontrar mais palavrões, mas à custa de muitos falsos positivos que penalizam a sua medida de precisão muitas vezes. No entanto, o método de Levenshtein mostra que consegue ser quase tão preciso quanto o método "Substring", ao passo que nenhum outro método consegue ascender aos níveis de cobertura elevados atingidos pelo reconhecedor "Levenshtein".

6.3.2 O Parâmetro de Tolerância

Iremos agora focar-nos nos métodos de reconhecimento "Substring" e "Levenshtein", que apresentam os resultados mais interessantes. Na Figura 9 encontra-se representado o impacto que os diferentes níveis de tolerância têm nos resultados, desta vez baseando a comparação na medida F1, onde se observa como estes dois métodos de reconhecimento se comportam de forma diferente. O eixo das ordenadas representa o valor do parâmetro de tolerância, que vai de muito tolerante (à esquerda) até completamente intolerante (à direita). Para ajudar a formar uma imagem mais completa, apresentam-se também a Figura 10, que mostra a contagem de falsos positivos, e a Figura 11, que apresenta a contagem de falsos negativos aquando do uso dos mesmos parâmetros.

O método "Substring" apresenta uma redução continuada dos falsos positivos, ou seja, uma melhoria da precisão, que acompanha o aumentar do grau de exigência. Porém, os falsos negativos, que se mantinham quase constantes com parâmetros inferiores a 0,5, aumentam quando se usam valores superiores. Ainda assim, o valor de F1 permanece algo estável, oscilando entre os 0,48 e os 0,60.

Por seu lado, o método "Levenshtein" revela resultados muito maus com valores de parâmetro muito permissivos. Produz 36 811 falsos positivos quando usado com o valor de parâmetro 0,1. No entanto estes valores decrescem de forma muito acelerada com o aumentar do valor do parâmetro, sem alguma vez ficarem abaixo do número de falsos positivos produzidos pelo método "Substring". A cobertura vai também reduzindo, mas a um ritmo muito inferior até ao valor de parâmetro 0,8, altura em que a medida F1 atinge o seu máximo de 0,64. Quando o método opera com valores de parâmetro 0,9 ou 1,0, difere pouco nos resultados do método "Substring".

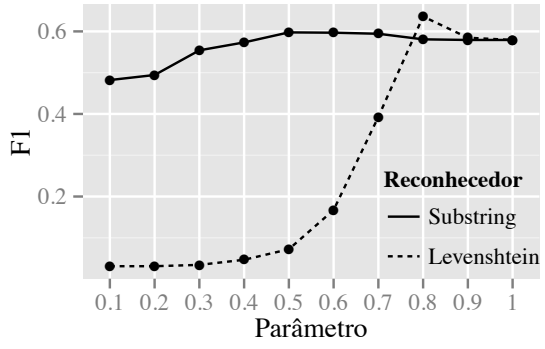

Figura 9: Efeito do nível de similaridade imposto aos métodos de reconhecimento na medida F1.

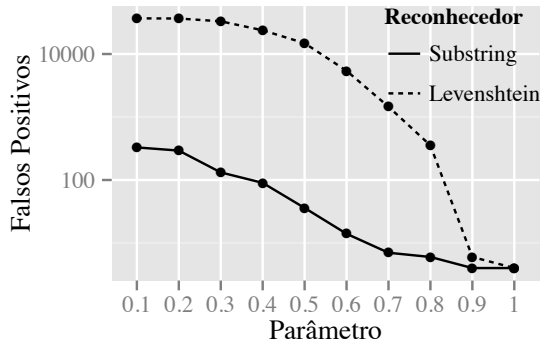

Figura 10: Contagem do número de falsos positivos obtidos com os métodos de reconhecimento "Substring" e "Levenshtein", quando usando diversos valores como parâmetro.

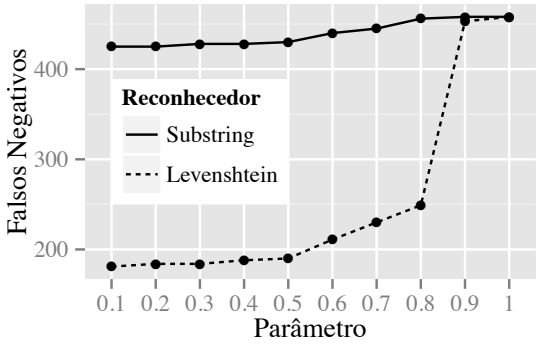

Figura 11: Número de falsos negativos gerados pelos métodos de reconhecimento "Substring" e "Levenshtein" em função do valor de parâmetro.

As medidas de 0,60 na precisão, 0,68 de cobertura e 0,64 de F1 deixam ainda uma margem significativa para melhoria, mesmo sendo este um teste muito exigente. Os objetivos de cobertura revelam-se muito difíceis de cumprir sem gerar uma torrente de falsos positivos, daí que se consideraram duas soluções possíveis. A primeira hipótese consiste no afrouxamento da avaliação de uma forma que permita ainda satisfazer as necessidades que propomos colmatar. A segunda recorre a uma fonte de dados externa, mais concretamente um dicionário, como forma de evitar muitos falsos positivos.

6.3.3 Reconhecimento mais tolerante

Se fôssemos menos rígidos na avaliação e em vez de exigir a variação exata do palavrão nos contentássemos em reconhecer qual dos 40 palavrões-base está correto, como é que os resultados melhorariam?

Na verdade, ao correr a experiência nesses parâmetros não se obtiveram ganhos significativos. Em vez dos 40 palavrões-base agruparam-se os palavrões em 22 *conceitos*, que abrangendo cada um 1 ou mais palavrões-base e suas variantes, mesmo que graficamente distintos.

Nestes moldes conseguiu-se apenas um ganho de 0,01 na medida F1 máxima, o que não é significativo. Isto poderá significar que os autores tendem a deixar indícios suficientes para compreender a variação correta do palavrão, e preferem ofuscar a parte mais comum do palavrão — que se compreende que seja a secção mais facilmente reconhecida por uma pessoa e consequentemente mais tolerante a alterações.

Visto que esta reformulação do problema não trouxe ganhos significativos, não foi mais perseguida.

6.3.4 Empregar um dicionário

Um dos problemas de aumentar a permissividade no processo de reconhecimento é que existem palavras graficamente semelhantes a palavrões mas que estão bem escritas. Por exemplo, "poder", "conta", "pilar" ou "nisso". O número de palavras cresce exponencialmente com o aumento do nível de tolerância que é permitida.

Por forma a ignorar todas as palavras bem escritas e que não são palavrões, optou-se por usar um dicionário, o que é uma decisão óbvia e não é nova (Sood, Antin e Churchill, 2012b).

Extraímos as 994 921 palavras do dicionário base de português do GNU/Aspell[8], do qual foram removidas as palavras do nosso léxico de palavrões. Repetiu-se depois a experiência, mas descartando imediatamente todos os átomos encontrados que constavam no dicionário, não as submetendo sequer ao reconhecimento de palavrões. Apresentamos apenas os valores com atomização "Simples" e normalização "Básica" que continuam a revelar-se a melhor aposta, analisando só o reconhecedor "Levenshtein".

Na Figura 12 pode ser visto o impacto que o dicionário teve na cobertura e na precisão, comparando os valores obtidos anteriormente sem o dicionário com os valores obtidos com o dicionário. Os pontos no extremo esquerdo (menor cobertura, maior precisão) representam os resultados obtidos com o mínimo de tolerância (parâmetro com valor 1.0). Observa-se que ambos os métodos convergem quando se exige maior similaridade das palavras, aproximando-se do comportamento do reconhecedor "Igual". A ligeira diferença nas medidas de cobertura pode ser explicada pela ofuscação de palavrões escrevendo-os como palavras no dicionário, como por exemplo a expressão "filho da pata", que já não chega a ser processada.

No outro extremo das linhas encontram-se os resultados gerados pelos parâmetros mais permissivos, onde é mais significativa a diferença de precisão conseguida, mas os níveis de cobertura são semelhantes. Isto acontece porque o efeito principal da filtragem de palavras é a eliminação de possíveis falsos positivos, enquanto o número de falsos negativos quase não é alterado.

O reflexo destes novos resultados na medida F1 está representado na Figura 13. Aqui comparam-se os resultados do reconhecedor "Levenshtein" quando usa e não usa o dicionário. É notória a melhoria dos valores de F1 para quase todos os níveis de exigência de similaridade, graças à presença deste filtro; mas 0,8 continua a ser o melhor valor de parâmetro. A partir desse valor, quando se dá pouca margem de manobra ao algoritmo de Levenshtein, usar ou não o dicionário faz pouca diferença.

[8] `http://aspell.net` visto em 2014-12-20

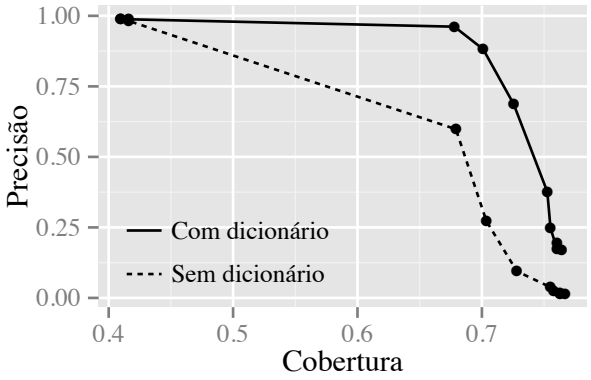

Figura 12: Precisão vs. Cobertura do método "Levenshtein" usando ou não um dicionário de português como filtro de palavras avaliadas.

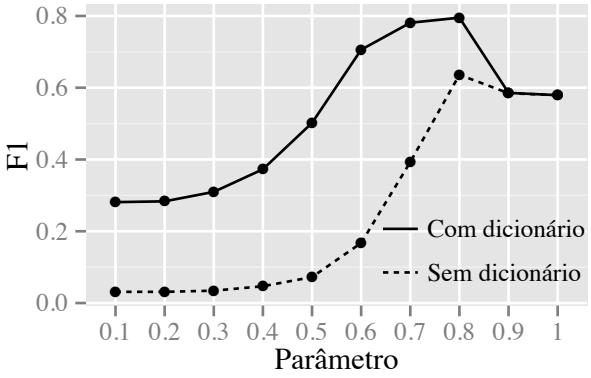

Figura 13: Impacto do dicionário na medida F1 para cada valor de parâmetro do reconhecedor "Levenshtein".

A conclusão que se pode tirar é que a exclusão das palavras presentes no dicionário permite que o reconhecedor opere com maior liberdade, sem no entanto se comprometer com falsos positivos. Assim é possível aumentar o nível de cobertura sem decréscimo significativo na precisão, como acontecia anteriormente. Os melhores resultados que foram obtidos são agora 0.96 de precisão, 0.68 de cobertura e 0.80 de F1.

7 Conclusão e trabalho futuro

O nosso estudo referente à desofuscação de palavrões dedicou-se principalmente a observar, medir e analisar. Observámos a prevalência do uso de palavrões no website SAPO Desporto, bem como os métodos usados para os disfarçar. Me-

abadalhocado	encornada	merdoso
badalhoca	encornadinhos	mijo
badalhoco	encornado	morcona
bardamerda	encornador	morconas
bastardos	encornados	morcão
bico	encornar	morcãozada
bicos	encornei	morcões
bosta	enraba	pachacha
bostas	enrabada	pachachinha
broche	enrabado	panasca
broches	enrabados	panascos
brochista	enrabar	paneleirice
cabrão	enrabá-lo	paneleiro
cabrões	enrrabar	paneleiros
cagadeira	esporra	paneleirote
cagado	esporrada	panisgas
cagalhão	foda	peida
cagalhões	foda-se	picha
caganeira	fodam	pila
cagarolas	fode	pilas
cago	fodei-vos	piroca
caguem	fodem	pirocas
caralhinhos	fodendo	pisso
caralho	foder	pizelo
caralhos	fodesses	pizelos
chulecos	fodeu	piça
cocó	fodida	piças
colhões	fodido	porcalhota
cona	fodidos	punhetas
conas	fodo	puta
corno	mamadas	putas
cornos	mamões	putinha
cornudas	maricas	putéfia
cornudo	mariconço	rabeta
cornudos	masturbar-se	rabetas
cu	merda	rameira
cuzinho	merditas	tomates

Tabela 6: Lista dos 111 palavrões anotados na coleção SAPO Desporto.

dimos o desempenho de vários métodos de pré-processamento de texto e reconhecimento de palavrões ofuscados, a fim de compreender qual é o contributo de cada um deles para o resultado final. Por fim, analisámos os valores obtidos e conseguimos melhorá-los significativamente.

Este estudo procurou essencialmente identificar os pontos fortes que convém manter na análise de palavrões, e os pontos fracos que necessitam de ser melhorados ou repensados por serem inadequados.

Começando pelo positivo, mostrámos que é possível usar métodos e técnicas comuns e frequentes para identificar e reconhecer palavrões, obtendo níveis de desempenho minimamente satisfatórios (0.96 de precisão, 0.68 de cobertura e 0.80 de F1). Foi também possível mostrar a importância do pré-processamento neste processo, designadamente a atomização e normalização, assim como comparar três formas de estabelecer correspondência entre os átomos no texto e os palavrões no léxico. Por fim, conseguiu-se medir o impacto resultante de uma filtragem dos átomos baseada num dicionário no processo de identificação e reconhecimento de palavrões.

A título de aspetos negativos, há a salientar todo o trabalho que resta ainda fazer para se poder atingir níveis de desempenho verdadeiramente bons nesta tarefa. Destacamos também o atomizador "Sylvester", por exemplo, que não materializou os resultados que se esperava obter do uso de uma ferramenta mais especializada, mesmo tendo em atenção que o estávamos a usar numa situação bastante específica e difícil para qualquer atomizador. Talvez exemplos de treino mais focados com texto ofuscado ajudassem a resolver parte do problema. A questão da ofuscação com espaços persistirá até que sejam desenvolvidos atomizadores com capacidade de aglutinação e que funcionem bem em ambientes ruidosos.

A normalização revelou-se muito útil, apesar de ter ainda aspetos passíveis de melhorar. Com uma língua como o inglês, por exemplo, a remoção dos acentos não teriam um impacto idêntico, e a tradução dos números para letras teria de ser bem pensada, visto que podem ser usados de forma mais fonética (e.g. "4ever" [forever], "l8er" [later], "2b or not 2b" [to be]). Como trabalho futuro propõe-se tomar em consideração o som das palavras, já que as substituições de letras tendem a manter a sua pronúncia (Laboreiro e Oliveira, 2014).

A distribuição estatística das letras pode também ser relevante para a normalização. Por exemplo, poucas palavras têm "k" em português, e por isso esta letra é propensa a ser removida do nosso alfabeto em prol de um "c" ou "q". Também é provável que na fase de normalização se consiga tratar da remoção de letras repetidas, permitindo que o processo de reconhecimento seja ainda mais simples por não ter de lidar com esta situação.

Ao nível dos reconhecedores, o algoritmo de Levenshtein revelou-se o mais versátil, conseguindo uma maior flexibilidade ao regular o equilíbrio entre cobertura (limite mais permissivo) e precisão (limite mais rígido) num espetro mais alargado que os outros dois. Ainda assim acreditamos que este algoritmo pode ser adaptado de forma a adequar-se melhor a esta tarefa. Uma possibilidade é modificar os pesos

das operações em função do caráter da palavra observado. Por exemplo, o custo de substituir um símbolo não alfabético por uma letra seria próximo de zero. Outra possibilidade é usar a coleção SAPO Desporto para treinar um mecanismo de desofuscação baseado em aprendizagem automática.

Outra linha de trabalho a estabelecer passa pela análise do contexto, que nós humanos usamos para nos ajudar a resolver as formas de ofuscação mais agressivas. Esta poderia basear-se na análise da frequência de n-gramas.

Agradecimentos

Este projeto teve o financiamento do CoLaboratório Internacional para Tecnologias Emergentes UT Austin | Portugal, projeto UTA-Est/MAI/0006/2009, e do SAPO Labs UP.

Referências

Almeida, José João. 2014. Dicionário aberto de calão e expressões idiomáticas, Outubro, 2014. `http://natura.di.uminho.pt/jjbin/dac`.

Cohen, Elliot D. 1998. Offensive message interceptor for computers, Agosto, 1998. Patent 5796948 A.

Constant, Noah, Christopher Davis, Christopher Potts, e Florian Schwarz. 2009. The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung: International Journal for Language Data Processing*, 33:5–21.

Jacob, Varghese, Ramayya Krishnan, Young U. Ryu, R. Chandrasekaran, e Sungchul Hong. 1999. Filtering objectionable internet content. Em *Proceedings of the 20th international conference on Information Systems*, ICIS '99, pp. 274–278, Atlanta, GA, USA. Association for Information Systems.

Jay, Timothy. 2009. The utility and ubiquity of taboo words. *Perspectives on Psychological Science*, 4(2):153–161.

Jay, Timothy e Kristin Janschewitz. 2007. Filling the emotional gap in linguistic theory: Commentary on Pot's expressive dimension. *Theoretical Linguistics*, 33:215–221.

Laboreiro, Gustavo e Eugénio Oliveira. 2014. What we can learn from looking at profanity. Em Jorge Baptista, Nuno Mamede, Sara Candeias, Ivandré Paraboni, Thiago A. S. Pardo, e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language*, volume 8775 of *Lecture Notes in Computer Science*. Springer International Publishing, pp. 108–113, Setembro, 2014. http://labs.sapo.pt/2014/05/obfuscation-dataset/.

Laboreiro, Gustavo, Luís Sarmento, Jorge Teixeira, e Eugénio Oliveira. 2010. Tokenizing micro-blogging messages using a text classification approach. Em *Proceedings of the fourth workshop on analytics for noisy unstructured text data*, AND '10, pp. 81–88, New York, NY, USA, Outubro, 2010. ACM. http://labs.sapo.pt/2011/11/sylvester-ugc-tokenizer/.

Levenshtein, V I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, Fevereiro, 1966.

Mahmud, Altaf, Kazi Zubair Ahmed, e Mumit Khan. 2008. Detecting flames and insults in text. Em *6th International Conference on Natural Language Processing (ICON-2008)*. Center for research on Bangla language processing (CRBLP), BRAC University.

Mehl, Matthias R. e James W. Pennebaker. 2003. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, 84(4):857–870.

Razavi, Amir Hossein, Diana Inkpen, Sasha Uritsky, e Stan Matwin. 2010. Offensive language detection using multi-level classification. Em Atefeh Farzindar e Vlado Keselj, editores, *Advances in Artificial Intelligence*, volume 6085 of *Lecture Notes in Computer Science*, pp. 16–27. Canadian Conference on Artificial Intelligence, Springer.

Sood, Sara Owsley, Judd Antin, e Elizabeth F. Churchill. 2012a. Profanity use in online communities. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pp. 1481–1490, New York, NY, USA. ACM.

Sood, Sara Owsley, Judd Antin, e Elizabeth F. Churchill. 2012b. Using crowdsourcing to improve profanity detection. Em *AAAI Spring Symposium: Wisdom of the Crowd*, volume SS-12-06 of *AAAI Technical Report*. AAAI.

Sood, Sara Owsley, Elizabeth F. Churchill, e Judd Antin. 2011. Automatic identification

of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, Outubro, 2011.

Spertus, Ellen. 1997. Smokey: Automatic recognition of hostile messages. Em *Proceedings of Innovative Applications of Artificial Intelligence (IAAI)*, pp. 1058–1065.

Thelwall, Mike. 2008. Fk yea I swear: cursing and gender in MySpace. *Corpora*, 3(1):83–107.

Wang, Wenbo, Lu Chen, Krishnaprasad Thirunarayan, e Amit P. Sheth. 2014. Cursing in English on Twitter. Em *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, Fevereiro, 2014.

Xiang, Guang, Bin Fan, Ling Wang, Jason Hong, e Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. Em *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1980–1984. ACM.

Xu, Zhi e Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. Em *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*.