Detecting Suicidal Ideation on Tweets

Vinícius de Carvalho Cardoso[1], **Dario Brito Calçada**[1,2]

[1]Universidade Estadual do Piauí (UESPI)
Campus Alexandre Alves de Oliveira – Parnaíba, PI – Brazil

[2]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos, SP – Brazil

{viniciuscc94,dariobcalcada}@gmail.com

Abstract. *According to the World Health Organization, every 40 seconds, one person dies of suicide in the world. Among young people aged from 15 to 29, suicide is the second leading cause of death. Still, these deaths can be prevented. In this scenario, social networks like Twitter can become real-time sources of information and help to prevent suicide. This paper presents an initial exploration of the problem of identifying individuals at risk of self-extermination in social networks that use Portuguese language. As a main scientific contribution, a set of tweet data, manually labeled by experts, was built and can be used for future research on the subject. As a preliminary evaluation, we applied machine learning algorithms for classification. The results indicate that the dataset can be used in a study to develop a real-time suicidal ideation tweet detection system.*

Resumo. *Segundo a Organização Mundial de Saúde, a cada 40s uma pessoa morre por suicídio no mundo. Entre jovens de 15 a 29, o suicídio é a segunda maior causa de mortes. Ainda assim, tais mortes podem ser prevenidas. Nesse cenário, redes sociais como o Twitter podem se tornar fontes de informação em tempo real e ajudar na prevenção do suicídio. O presente trabalho faz uma exploração inicial ao problema de identificar indivíduos em risco de suicídio nas redes sociais em Língua Portuguesa. Como principal contribuição científica, foi construído um dataset de tweets, manualmente rotulado por especialistas, que pode ser usado em futuras pesquisas acerca do tema. Para validação do conjunto de dados, foram utilizados algoritmos de aprendizado de máquina para classificação, sendo o resultado promissor, indicando que o dataset pode ser utilizado em pesquisas para elaboração de um sistema de detecção de tweets de ideação suicida em tempo real.*

1. Introdução

Segundo a OMS (Organização Mundial da Saúde), a cada 40s uma pessoa morre por suicídio no mundo. É estimado que, só em 2012, 804.000 suicídios tenham ocorrido, representando uma média anual global de 11,4 a cada 100.000 pessoas. Entre jovens de 15 a 29 anos, o suicídio é a segunda maior causa de mortes globalmente. Porém, esses números podem ser ainda maiores, uma vez que, em muitas ocasiões, devido a uma classificação equivocada como acidente ou outra causa de morte, ou, ainda, em países onde o ato é ilegal, casos de suicídio não são reportados [OMS 2014].

Apesar desses números preocupantes, mortes por suicídio podem ser prevenidas [Bailey et al. 2011]. Nesse cenário, a detecção precoce de risco de suicídio pode prover a base para programas de intervenção, o que pode ser efetivo em prevenir tais mortes. E, embora pessoas suicidas não se sintam motivadas a revelar seus pensamentos ou planos antes de uma tentativa, o período que precede um suicídio pode conter indícios a respeito da intenção do indivíduo [OMS 2014].

Diante de uma realidade de indícios, métodos tradicionais de prevenção não conseguem identificar pessoas em risco de suicídio em tempo real [McCarthy 2010]. Nesse cenário, redes sociais a exemplo do *Twitter*, que é uma plataforma de expressão pessoal, podem se tornar fontes de informação em tempo real e ajudar na prevenção do suicídio [Jashinsky et al. 2014]. De fato, pesquisas têm mostrado que indivíduos em risco estão recorrendo a tecnologias contemporâneas (fóruns, *micro-blogs*) para expressar seus problemas sem precisar encarar alguém pessoalmente [Moreno et al. 2011, De Choudhury et al. 2013]. Inclusive, casos de vítimas de suicídio escrevendo seus pensamentos finais nessas comunidades *online* já foram reportados [Gunn and Lester 2015, Kailasam and Samuels 2015]. Não obstante, muitos trabalhos têm encontrado relação entre risco de suicídio e padrões linguísticos em *posts* das redes sociais [McCarthy 2010, Sueki 2015].

Apesar de não haver um consenso entre diferentes comunidades de pesquisa [Miner et al. 2012], a Mineração de Textos (MT) pode ser vista como a aplicação de um conjunto de técnicas usadas para analisar dados não estruturados e descobrir padrões que não eram conhecidos previamente [Aggarwal and Zhai 2012]. Assim, a MT pode ser tratada como uma especialização da Mineração de Dados. Enquanto a Mineração de Dados trata os dados estruturados, a MT lida com textos escritos em língua natural (dados não estruturados). Com o crescente aumento e variedade de documentos textuais, tanto em redes sociais e Web em geral quanto internamente em organizações, as técnicas de MT têm se tornado essenciais no apoio à descoberta de conhecimento. Com isso, as fontes de textos, bem como as aplicações da MT, são variadas.

De forma geral, o processo de MT pode ser visto como um processo formado por cinco etapas, conforme ilustrado na Figura 1. Esse processo se inicia com a especificação de seus objetivos na etapa de Identificação do Problema. Nesta etapa, o analista, especialista em MT, deve delimitar o escopo da mineração, preferivelmente trabalhando com um especialista do domínio de aplicação. Devem ser definidas as coleções de textos que serão mineradas e como os resultados serão utilizados. As especificações definidas na etapa de Identificação do Problema guiarão as próximas etapas do processo de MT, as quais podem ser executadas em ciclos de preparação dos dados (etapa de Pré-processamento), descoberta de conhecimento (etapa de Extração de Padrões) e avaliação do conhecimento (etapa de Pós-processamento) [Sinoara et al. 2017].

Visando reduzir o número de termos e amenizar os problemas da alta dimensionalidade e esparsidade, pode-se utilizar algumas técnicas de pré-processamento, como:

- Remoção de *stopwords*: a remoção de *stopwords* visa a eliminação de palavras que não trazem informação relevante para o processo de MT. Essas palavras, chamadas de *stopwords*, normalmente são palavras que possuem as funções de artigos, preposições, pronomes e conjunções. No entanto, também podem ser identificadas *stopwords* específicas do domínio de aplicação do processo, ou seja, palavras que

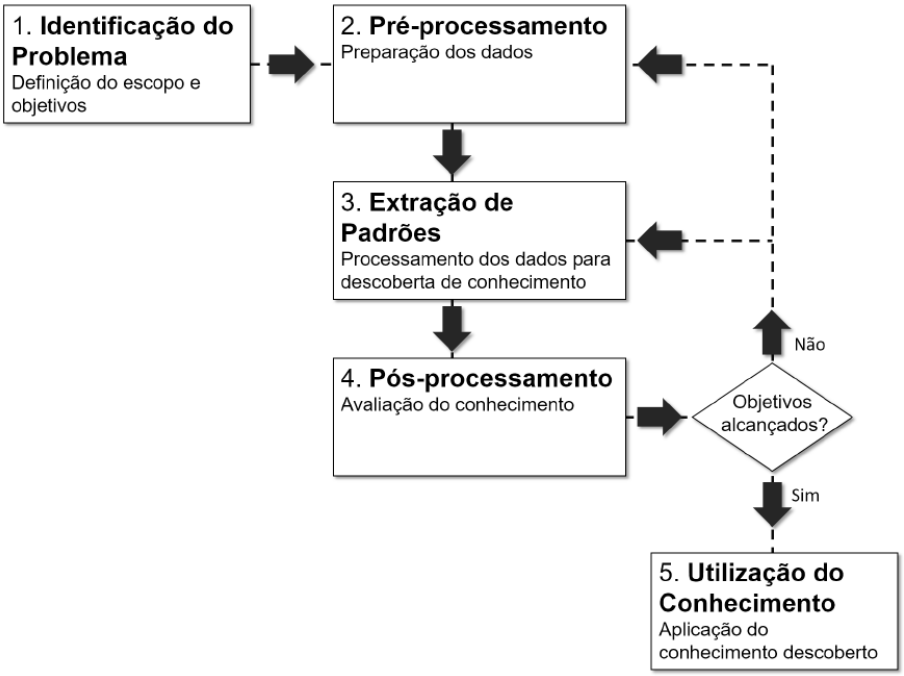

Figura 1. Processos de Mineração de Textos. Adaptado de [Sinoara et al. 2017]

sabidamente são frequentes na coleção e que não distinguem classes ou grupos que espera-se identificar com a Mineração de Textos.

- Normalização: a normalização dos textos visa eliminar as diversas variações que as palavras podem sofrer, como por exemplo variações de gênero e número dos substantivos e conjugações dos verbos. A normalização pode ser realizada por meio de três técnicas: (i) radicalização (*stemming*), que reduz cada palavra ao seu radical (ou palavra raiz); (ii) lematização, que reduz cada palavra a seu lema (ou forma canônica), ou seja, verbos são reduzidos ao infinitivo, e substantivos e adjetivos são reduzidos a forma masculina singular; e (iii) substantivação, que transforma a palavra para que ela tenha o comportamento sintático/semântico semelhante a de um substantivo.

A busca pela detecção de ideação suicida em tempo real se apresenta como uma alternativa viável no auxílio a órgãos que cuidam deste tipo de autoextermínio. Este estudo visa preencher as lacunas na literatura em relação à análise e à classificação de conteúdo relacionado ao suicídio em Língua Portuguesa. Os principais objetivos deste trabalho são: i) criar um *dataset* público composto por *tweets* escritos em português para futuras pesquisas acerca do tema, e ii) comparar o desempenho de diferentes algoritmos de classificação de textos na distinção de *tweets* com e sem ideação suicida.

Os resultados foram promissores já que foi construído um *dataset* bem amplo de *tweets* classificados como sendo de ideação suicida ou não. O *dataset* foi rotulado por especialistas na área que validaram as respostas formando um conjunto de dados robusto e que poderá ser utilizado em outras pesquisas. Os algoritmos de classificação supervisionados apresentaram bom comportamento diante da complexidade dos dados apresentados.

Além dessa Seção introdutória o presente artigo é organizado de modo que na

Seção 2 são apresentados as pesquisas que inspiraram este trabalho, bem como lacunas que ainda não haviam sido elucidadas. Na Seção 3 são apresentados os métodos utilizados para a coleta dos *tweets* e a rotulação dos dados, bem como os algoritmos utilizados na tarefa de classificação. Os resultados obtidos são apresentados na Seção 4. Por fim, na Seção 5 são apresentadas as considerações finais sobre a pesquisa, além de possibilidades de trabalhos futuros.

2. Trabalhos Relacionados

Nos últimos anos, diversos trabalhos investigaram a possibilidade de criar modelos para a classificação de conteúdo relacionado ao suicídio nas redes sociais. Em [Abboute et al. 2014], foi descrito um processo completo para, automaticamente, coletar *tweets* suspeitos de acordo com um vocabulário de termos, criado pelos próprios autores, que pessoas suicidas costumam utilizar. Com a obtenção de um *corpus* que também incluía casos comprovados, os *tweets* foram classificados em "*risky*" e "*non risky*". Então o desempenho de seis classificadores foram comparados, obtendo uma acurácia de aproximadamente 64% para o algoritmo *Naïve Bayes*.

Focados principalmente na criação do corpus, [Desmet and Hoste 2014] introduziram uma estratégia de anotação em cascata para reconhecer conteúdo suicida. Em tal esquema, os textos, obtidos do *Netlog* (rede social particularmente popular entre jovens na Holanda), foram analisados por aspectos como relevância, origem, sujeito, severidade e, ainda, se apresentavam fatores protetivos ou de risco para suicídio. Como melhor resultado obtido durante os experimentos, atingiu-se uma precisão de 79% com o algoritmo SVM (*Support Vector Machines* - Máquinas de Vetores de Suporte).

Dividindo os *tweets* em três classes, ao invés de apenas duas, [O'Dea et al. 2015] tinham como objetivo examinar se o nível de "preocupação" para um *tweet* relacionado ao suicídio poderia ser determinado baseando-se somente em seu conteúdo. Como principal resultado obtido, o classificador SVM classificou corretamente 80% dos *tweets* rotulados como "strongly concerning", i.e., extremamente preocupante, o que indica uma possível mensagem de uma pessoa com ideação suicida.

Em seu trabalho, [Burnap et al. 2017] foram mais a fundo. Primeiro, eles dividiram os *tweets* em mais classes (sete ao todo) para distinguir entre conteúdo mais preocupante (ex. ideação suicida) e outros também relacionados ao suicídio (ex. relatos de suicídio, campanhas preventivas, etc.). Além disso, na etapa de experimentos, os autores não só avaliaram o desempenho dos algoritmos individualmente, mas também construíram um *ensemble* desses classificadores, o qual levou aos melhores resultados obtidos.

Até a data de escrita deste artigo, nenhum trabalho com a mesma proposta dos acima citados foi encontrado em Língua Portuguesa. Com o intuito de preencher esta lacuna em nossa literatura, este trabalho visa realizar um estudo inicial na identificação de *tweets*, escritos em Português, que apresentam ideação suicida.

3. Materiais e Métodos

O trabalho está dividido em três principais etapas: i) coleta de dados, ii) rotulação manual por especialistas e iii) classificação com o uso de algoritmos de aprendizado de máquina.

3.1. Coleta de Dados

Como primeira etapa do processo de Mineração de Textos deve-se obter um conjunto de documentos pertencentes ao domínio do problema a ser resolvido. Assim, de 15 de março a 15 de abril, utilizou-se de uma API[1] (*Application Programming Interface* - Interface de Programação de Aplicativos), disponibilizada pelo próprio *Twitter*, para automaticamente coletar um corpus de *tweets* que continham termos ou expressões próprios do vernáculo da ideação suicida [O'Dea et al. 2015]. Tais palavras e frases, por terem sido definidas em um trabalho em Língua Inglesa (ex. *suicide, suicidal, end my life, not worth living*), foram primeiramente traduzidas para o Português antes da realização da coleta. Na Tabela 1 são listadas todas as palavras e expressões selecionadas para captação dos *tweets*.

Tabela 1. Palavras e expressões do vernáculo da ideação suicida.

Original	Tradução
Suicidal	Suicida
Suicide	Suicídio
Kill myself	Me matar
My suicide note	Meu bilhete suicida
My suicide letter	Minha carta de suicídio
End my life	Acabar com a minha vida
Never wake up	Nunca acordar
Can't go on	Não consigo continuar
Not worth living	Não vale a pena viver
Ready to jump	Pronto para pular; pronto pra pular
Sleep forever; go to sleep forever	Dormir para sempre; dormir pra sempre
Want to die	Quero morrer
Be dead	Estar morto
Better off without me	Melhor sem mim
Better off dead	Melhor morto
Suicide plan	Plano de suicídio
Tired of living	Cansado de viver
Die alone	Morrer sozinho
Don't want to be here	Não quero estar aqui

Com relação às palavras e expressões utilizadas para recuperar os *tweets* neste trabalho, algumas considerações devem ser feitas. Primeiro que, diferentemente do Inglês, os adjetivos sofrem flexão de gênero no Português. Optou-se, então, nos casos onde a expressão original tivesse adjetivo que o mesmo seria traduzido apenas para o masculino (ex. *tired of living* - cansado de viver). Outra decisão tomada durante essa tarefa foi, levando em consideração a natureza informal da escrita de *tweets*, traduzir alguns casos considerando tanto a norma padrão da língua quanto a informal (ex. *sleep forever* - (dormir para sempre; dormir pra sempre)).

Embora a API do *Twitter* permita coletar vários dados a respeito de um *tweet* como o nome do usuário que o publicou, o número de *retweets*, local e data da publicação, etc.,

[1] https://developer.twitter.com/en/docs/tweets/search/overview

por questões que envolvem os objetivos deste trabalho, além das questões éticas, foram coletados e armazenados apenas o conteúdo textual dos *tweets*.

3.2. Rotulação por Especialistas

Para essa fase, uma amostra de 1190 *tweets* (aproximadamente 25% dos 4732 coletados) foi escolhida para anotação manual. Para a realização dessa tarefa, uma equipe de especialistas, composta por três psicólogos, foi formada. Os *tweets* foram organizados e tabulados a fim de que cada um dos especialistas pudesse realizar a leitura completa dos mesmos e responder a pergunta: "Esse *tweet* apresenta ideação suicida?".

Os *tweets* da amostra apresentada puderam ser classificados em uma de três categorias possíveis: "positivo", "negativo" e "indefinido". A fim de ser rotulado como "positivo", o *tweet* deveria indicar ideação suicida por parte do usuário, o que poderia ser expresso por meio da presença de fatores de risco no texto. Desse modo, sempre que um *tweet* indicasse depressão ou outras desordens psicológicas, tentativas anteriores de suicídio, violência familiar, bullying, sentimentos de isolamento, impulsividade, entre outros fatores (*American Foundation for Suicide Prevention* [AFSP 2013]), receberia o rótulo "positivo".

Por outro lado, se um *tweet* apresentasse sarcasmo, fizesse referência a notícias ou, ainda, a campanhas de prevenção ao suicídio, o rótulo "negativo" deveria ser escolhido. Por fim, caso o especialista não conseguisse classificar o *tweet* como positivo ou negativo, ele poderia atribuir-lhe a categoria "indefinido".

Após a classificação manual realizada pelos 3 (três) especialistas, na qual cada um rotulou todos os *tweets* da amostra, sem nenhum tipo de contato com os outros dois, foi realizado um processo para selecionar os *tweets* que comporiam o *dataset* final a ser utilizado na etapa posterior.

Com a intenção de construir um *dataset* apenas com os *tweets* classificados como "positivo" e "negativo", todos os *tweets* rotulados de forma unânime como "indefinido" pela equipe de especialistas foram descartados. Todos os *tweets* que não receberam o mesmo rótulo por parte dos três especialistas também foram excluídos. O *dataset* final é composto por 699 instâncias e está disponível on-line no link `http://bit.ly/2SioqW7`.

3.3. Classificação utilizando algoritmos de aprendizado supervisionado

Com os dados devidamente rotulados pela equipe de especialistas, algoritmos de aprendizado de máquina foram aplicados para derivar um classificador de texto capaz de predizer automaticamente a classe de cada *tweet*. A API do WEKA[2] (*Waikato Environment for Knowledge Analysis*) foi utilizada para comparar o desempenho de um número de classificadores.

No entanto, como se trata da categorização de textos, técnicas de PLN (Processamento de Linguagem Natural) foram aplicadas, principalmente na etapa de pré-processamento dos *tweets*. A principal atividade realizada na etapa de Pré-processamento é a representação dos textos em um formato aceito pelo algoritmo a ser utilizado na

[2]http://www.cs.waikato.ac.nz/ml/weka/

Extração de Padrões. Os algoritmos tradicionais de Aprendizado de Máquina, que normalmente são utilizados na mineração de dados estruturados, assumem que os dados são apresentados em um formato conhecido como matriz atributo-valor. Nessa matriz, cada instância (ou exemplo) corresponde a uma linha e seus atributos (características que descrevem esse exemplo) correspondem às colunas.

Nos processos de MT, após a identificação do problema (Figura 1), é necessário o pré-processamento da informação. Nesta etapa busca-se realizar a mudança de um dado não estruturado (texto) para um estruturado. Desse modo, passos como *tokenization*, remoção de pontuação, remoção de *stopwords* e aplicação de *stemming*, comuns a qualquer pré-processamento de texto tradicional, foram realizados. Os *tweets* foram representados com a técnica de *bag-of-words* com a frequência simples de cada *token* presente.

Adicionalmente, outras técnicas especificamente aplicadas a *tweets* como remoção de referências a usuários e remoção de URLs, foram utilizadas [Go et al. 2009]. Assim, cada *tweet* foi transformado em vetor de *n-grams* (nesse caso, *unigrams*, ou seja cada palavra foi considerada como uma *feature*), seguindo a abordagem para representação de textos *bag-of-words*.

Neste trabalho foram aplicados sete algoritmos de classificação indutiva, tradicionais e estado da arte, sendo cinco algoritmos da biblioteca Weka [Witten and Frank 2005] e dois algoritmos baseados em redes bipartidas [Rossi et al. 2014, Rossi et al. 2016]. Apesar das representações avaliadas neste trabalho serem representações no modelo espaço-vetorial, a utilização de algoritmos baseados em redes bipartidas também é possível. As redes bipartidas podem ser obtidas por mapeamento direto de representações documento-termo, ou seja, representações no modelo espaço-vetorial [Rossi et al. 2016].

Com o objetivo de se avaliar as representações em cenários bem diversificados, buscou-se diferentes variações dos parâmetros utilizados para cada algoritmo. Os valores selecionados para os parâmetros dos algoritmos foram baseados nos valores utilizados na extensa avaliação experimental de classificação indutiva supervisionada realizada por [Rossi et al. 2016]. Assim, a seleção de parâmetros foi realizada visando a diversificação dos modelos e o uso de configurações utilizadas em outros trabalhos da literatura e que apresentam bons resultados. Os algoritmos e os parâmetros utilizados nas avaliações experimentais deste trabalho são apresentados a seguir.

- Naive Bayes (NB)
  O algoritmo Naive Bayes é um classificador probabilístico baseado no "Teorema de Bayes", o qual foi criado por Thomas Bayes (1701 - 1761). Atualmente, o algoritmo se tornou popular na área de Aprendizado de Máquina para categorizar textos baseado na frequência das palavras usadas, podendo ser utilizado para identificação do assunto ao qual o texto se refere.
- Multinomial Naive Bayes (MNB)
  O algoritmo MNB implementa o algoritmo Naive Bayes para dados distribuídos multinomialmente e é uma das duas variantes clássicas do NB usadas na classificação de texto (onde os dados são tipicamente representados como contagens vetoriais de palavras).
- C4.5

Foi utilizada a implementação da ferramenta Weka (J48). Foram utilizados os níveis de confiança 0,15, 0,20 e 0,25.

- Support Vector Machine (SVM)

  Foi utilizado o algoritmo *Sequential Minimal Optimization* (SMO), implementação da ferramenta Weka. Foram utilizados três tipos de kernel: linear, polynomial (com expoente = 2) e radial basis function. Os valores de C utilizados para cada tipo de kernel foram: 0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4} e 10^{5}.

- k-Nearest Neighbors (k-NN)

  Foi utilizada a implementação da ferramenta Weka (IBk). O algoritmo foi utilizado com e sem o voto ponderado pela distância entre os exemplos. Foram aplicadas duas medidas de distância: Cosseno e Euclidiana. Os valores de k utilizados foram: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45 e 55.

- $IMBHN^{C}$

  Versão do algoritmo Inductive Model based on Bipartite Heterogeneous Networks, algoritmo baseado em redes heterogêneas bipartidas. Essa versão do algoritmo altera as informações de classe dos termos caso as informações de classe correntes produzam um erro de classificação. Foram utilizadas as taxas de correção de erro de 0,01, 0,05, 0,1 e 0,5. O número máximo de iterações foi definido em 1000 e o critério de parada foi definido como sendo o limiar de 0,01 para o erro quadrático médio.

- $IMBHN^{R}$

  Versão do algoritmo Inductive Model based on Bipartite Heterogeneous Networks, algoritmo baseado em redes heterogêneas bipartidas. Essa versão do algoritmo realiza uma regressão para induzir as informações de classe dos termos. Foram utilizadas as taxas de correção de erro de 0,01, 0,05, 0,1 e 0,5. O número máximo de iterações foi definido em 1000 e o critério de parada foi definido como sendo o limiar de 0,01 para o erro quadrático médio.

Considerando-se o uso de algoritmos de aprendizado indutivo supervisionados, o problema de classificação automática de textos é definido como se segue. Dados um conjunto de classes (C) e uma coleção de documentos rotulados (D), documentos cuja classe é conhecida, um algoritmo indutivo supervisionado induz uma função F que mapeia os documentos de D a classes de C ($F : D \rightarrow C$). A função F é chamada de modelo de classificação (ou classificador) e é utilizada para predizer a classe de novos documentos. Esse processo é ilustrado na Figura 2.

A performance dos classificadores foram avaliadas por meio das medidas Acurácia (equivalente a Micro-F1), Precisão e *Recall*. As medidas foram obtidas por meio de *10-fold cross-validation*. Todos os algoritmos foram avaliados segundo as mesmas partições dos dados no processo de treinamento e teste executados. As avaliações experimentais foram executadas por meio da ferramenta Text Categorization[3], disponibilizada por Rossi e colabotadores [Rossi et al. 2016].

[3]Text Categorization tool: `http://sites.labic.icmc.usp.br/ragero/thesis/text_categorization_tool`

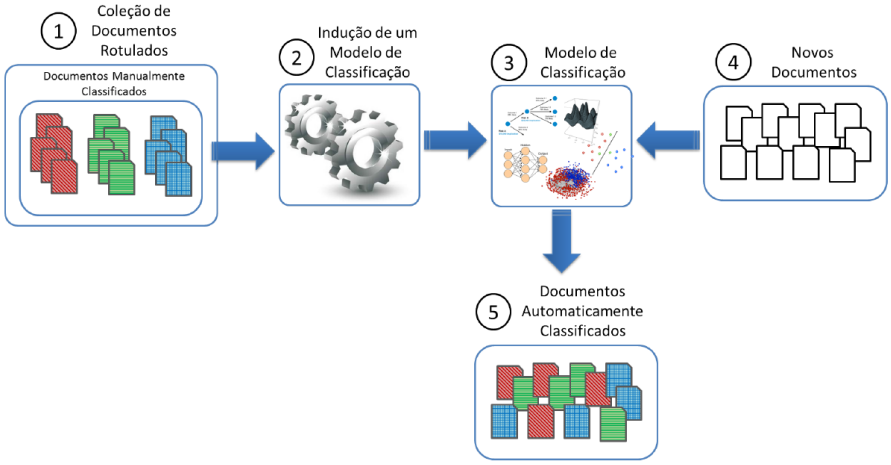

Figura 2. Esquema ilustrativo da classificação automática de textos por meio de aprendizado indutivo supervisionado. [Rossi 2016]

4. Resultados e Discussão

Um dos objetivos deste trabalho foi a construção de um *dataset* rotulado a ser utilizado em estudos acerca da detecção de ideação suicida em *tweets*. Assim, na seção 4.1, será descrito o processo de construção de tal *dataset*. Já na seção 4.2, é feita a comparação entre os resultados obtidos por cada algoritmo de Aprendizado de Máquina na tarefa de classificação de *tweets*.

4.1. Dataset

O *dataset* construído para este trabalho é composto por 699 instâncias. Cada instância representa um *tweet* rotulado manualmente e de forma unânime pela equipe de especialistas em "positivo" (211 instâncias) ou "negativo" (488 instâncias). Para ser classificado como "positivo", o *tweet* deveria apresentar indícios de ideação suicida por parte do usuário. Por outro lado, sempre que um *tweet* fizesse referência ao suicídio de terceiros, campanhas de prevenção ao suicídio, ou apresentasse sarcasmo e ironia, ele deveria ser classificado como "negativo".

Por fim, com os *tweets* rotulados, foi realizado o pré-processamento dos mesmos. Assim, etapas como *tokenization*, remoção de pontuação e caracteres especiais, remoção de *stopwords* e aplicação de *stemming*, usualmente presentes na Mineração de Textos tradicional, foram aplicadas. Adicionalmente, devido a características próprias dos *tweets*, como tamanho curto e o uso de linguagem informal por parte dos usuários, passos como remoção de referências a outros usuários e URLs, assim como a expansão de abreviaturas comuns na Internet foram realizados. Após todas essas etapas, obteve-se a configuração final do *dataset*, disponível on-line em `http://bit.ly/32aDPf9`.

4.2. Classificação automática dos *tweets*

Os experimentos de classificação foram realizados para validar o conjunto gerado. Na Tabela 2 são apresentados os melhores valores para acurácia obtidos para cada algoritmo em todas as configurações descritas na Seção anterior, juntamente com seus respectivos valores para Precisão e *Recall*. Observa-se que os desempenhos variaram bastante para

todas as medidas. Para Acurácia, os valores vão de 30,18%, obtido pelo algoritmo SVM em sua configuração Linear - 0, até 81,83% pelo mesmo algoritmo em sua configuração SVM (SMO) - RBFKernel - 10^1, representando que o conjunto de dados permite uma sensibilização dos algoritmos de classificação. Portanto, o conjunto de dados gerado pode ser utilizado para elaboração de sistemas que utilizam técnicas de aprendizado de máquina a fim de detecção de *tweets* de ideação suicida em tempo real.

Comportamento similar é observado em relação às medidas de Precisão e *Recall*, nos quais os valores obtidos encontram-se no intervalo de 0,15 a 0,81 para Precisão e de 0,5 a 0,79 para Recall, mas sempre para configurações do algoritmo SVM (SMO), corroborando com o observado na literatura como o algoritmo que apresenta os melhores resultados para a tarefa de classificação de textos, principalmente quando a abordagem *bag-of-words* para a representação dos textos é usada.

Nota-se também que os resultados alcançados neste artigo superam aqueles obtidos nos trabalhos relacionados citados anteriormente. Deve-se levar em consideração, obviamente, que tais trabalhos foram realizados para Língua Inglesa e sob configurações experimentais diferentes em relação a este. Ademais, por se tratar do primeiro estudo a abordar tal tema para textos em português, os resultados aqui obtidos poderão servir de *baseline* para futuros trabalhos.

5. Conclusões

Neste artigo, foi apresentado todo o processo de construção de um *dataset* formado com *tweets* de ideação suicida escritos em Língua Portuguesa. A não existência de um conjunto de dados para estudo desta temática em pesquisas de PLN motivou a execução deste trabalho, bem como pelo fato do número crescente de suicídios, que ocasiona um problema de saúde pública em vários países.

Para a construção do dataset, um corpus composto por *tweets* contendo palavras ou expressões condizentes com o vocabulário suicida foi coletado. Uma equipe de especialistas então foi encarregada de classificar manualmente os *tweets*, baseando-se apenas em seu conteúdo. No fim, após a realização do pré-processamento dos textos, obteve-se um *dataset* rotulado composto por *tweets* escritos em português e que está disponibilizado publicamente para futuras pesquisas no tema. A construção e a disponibilidade do *dataset* é a principal contribuição científica desta pesquisa.

Para validação do uso do *dataset*, foram comparados os desempenhos de um número de algoritmos de classificação para a tarefa de identificar conteúdo relacionado à ideação suicida no *Twitter*. A quantidade de algoritmos de classificação utilizados em variadas configurações possibilitou observar o comportamento destes algoritmos diante de um conjunto de dados coletado automaticamente do *Twitter* e rotulados manualmente por especialistas. A convergência de todos os algoritmos possibilitou a validação do uso do *dataset* construído, o que pode alavancar pesquisas nessa temática dentro da área de Mineração de Textos.

Como trabalho futuro, poderá se utilizar *ensembles* de classificadores para medir o poder preditivo dos modelos de classificação. Outra melhoria possível seria a utilização de *features* que expressem as características emotivas e estruturais dos *tweets*. Por fim, outras abordagens, além da *bag-of-words*, para a representação dos textos também poderiam ser exploradas.

Acknowledgment

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Agradecimentos também aos especialistas psicólogos que participaram desta pesquisa e de todos aqueles que trabalham em prol da prevenção ao suicídio.

Referências

Abboute, A., Boudjeriou, Y., Entringer, G., Azé, J., Bringay, S., and Poncelet, P. (2014). Mining twitter for suicide prevention. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 250–253. Springer.

AFSP (2013). *Risk factors and warning signs*. American Foundation for Suicide Prevention.

Aggarwal, C. C. and Zhai, C. (2012). *Mining Text Data*. Springer, 1 edition.

Bailey, R. K., Patel, T. C., Avenido, J., Patel, M., Jaleel, M., Barker, N. C., Khan, J. A., All, S., and Jabeen, S. (2011). Suicide: current trends. *Journal of the National Medical Association*, 103(7):614–617.

Burnap, P., Colombo, G., Amery, R., Hodorog, A., and Scourfield, J. (2017). Multi-class machine classification of suicide-related communication on twitter. *Online social networks and media*, 2:32–44.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Desmet, B. and Hoste, V. (2014). Recognising suicidal messages in dutch social media. In *9th international conference on language resources and evaluation (LREC)*, pages 830–835.

Go, A., Huang, L., and Bhayani, R. (2009). Twitter sentiment analysis. *Entropy*, 17:252.

Gunn, J. F. and Lester, D. (2015). Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, 17(3).

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., and Argyle, T. (2014). Tracking suicide risk factors through twitter in the us. *Crisis*.

Kailasam, V. and Samuels, E. (2015). Can social media help mental health practitioners prevent suicides? *Current Psychiatry*, 14(2):37–51.

McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of affective disorders*, 122(3):277–279.

Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., and Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Academic Press, 1 edition.

Moreno, M. A., Jelenchick, L. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E., and Becker, T. (2011). Feeling bad on facebook: depression disclosures by college students on a social networking site. *Depression and anxiety*, 28(6):447–455.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., and Christensen, H. (2015). Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

OMS (2014). *Preventing suicide: A global imperative*. World Health Organization.

Rossi, R. G. (2016). *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*. PhD thesis, Universidade de São Paulo.

Rossi, R. G., Lopes, A. A., and Rezende, S. O. (2016). Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts. *Information Processing and Management*, 52(2):217–257.

Rossi, R. G., Lopes, A. d. A., Faleiros, T. d. P., and Rezende, S. O. (2014). Inductive model generation for text classification using a bipartite heterogeneous network. *Journal of Computer Science and Technology*, 29(3):361–375.

Sinoara, R. A., Antunes, J., and Rezende, S. O. (2017). Text mining and semantics: a systematic mapping study. *Journal of the Brazilian Computer Society*, 23(9):1–20.

Sueki, H. (2015). The association of suicide-related twitter use with suicidal behaviour: a cross-sectional study of young internet users in japan. *Journal of affective disorders*, 170:155–160.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2 edition.

Tabela 2. Resultados experimentais para Acurácia, Precisão e Recall.

Algoritmo	Acurácia(%)	Precisão	Recall
NB	80,10766046	0,763979679	0,741911156
MNB	72,95445135	0,679365513	0,692273321
J48 - 0,15	77,39130435	0,768066501	0,679624361
J48 - 0,20	77,67701863	0,749121268	0,718433795
J48 - 0,25	78,39544513	0,751050408	0,737705167
SVM (SMO) - Linear - 10^{-3}	69,81780538	0,349089027	0,5
SVM (SMO) - Linear - 10^{-2}	69,81780538	0,349089027	0,5
SVM (SMO) - Linear - 0	30,18219462	0,150910973	0,5
SVM (SMO) - Linear - 10^{-1}	80,97101449	0,793297563	0,728630692
SVM (SMO) - Linear - 10^{0}	79,68944099	0,762653642	0,793762378
SVM (SMO) - Linear - 10^{-4}	69,81780538	0,349089027	0,5
SVM (SMO) - Linear - 10^{-5}	69,81780538	0,349089027	0,5
SVM (SMO) - Linear - 10^{1}	76,39337474	0,732099093	0,767420035
SVM (SMO) - Linear - 10^{2}	76,39337474	0,732099093	0,767420035
SVM (SMO) - Linear - 10^{3}	76,39337474	0,732099093	0,767420035
SVM (SMO) - Linear - 10^{4}	76,39337474	0,732099093	0,767420035
SVM (SMO) - Linear - 10^{5}	76,39337474	0,732099093	0,767420035
SVM (SMO) - PolyKernel - 10^{-3}	69,81780538	0,349089027	0,5
SVM (SMO) - PolyKernel - 10^{-2}	74,82194617	0,789572188	0,590168234
SVM (SMO) - PolyKernel - 0	30,18219462	0,150910973	0,5
SVM (SMO) - PolyKernel - 10^{-1}	79,26501035	0,76097431	0,79299678
SVM (SMO) - PolyKernel - 10^{0}	77,54865424	0,748029147	0,788019517
SVM (SMO) - PolyKernel - 10^{-4}	69,81780538	0,349089027	0,5
SVM (SMO) - PolyKernel - 10^{-5}	69,81780538	0,349089027	0,5
SVM (SMO) - PolyKernel - 10^{1}	77,12008282	0,745385415	0,785169973
SVM (SMO) - PolyKernel - 10^{2}	77,12008282	0,745385415	0,785169973
SVM (SMO) - PolyKernel - 10^{3}	77,12008282	0,745385415	0,785169973
SVM (SMO) - PolyKernel - 10^{4}	77,12008282	0,745385415	0,785169973
SVM (SMO) - PolyKernel - 10^{5}	77,12008282	0,745385415	0,785169973
SVM (SMO) - RBFKernel - 10^{-3}	69,81780538	0,349089027	0,5
SVM (SMO) - RBFKernel - 10^{-2}	69,81780538	0,349089027	0,5
SVM (SMO) - RBFKernel - 0	30,18219462	0,150910973	0,5
SVM (SMO) - RBFKernel - 10^{-1}	69,81780538	0,349089027	0,5
SVM (SMO) - RBFKernel - 10^{0}	73,10766046	0,811008185	0,556079371
SVM (SMO) - RBFKernel - 10^{-4}	69,81780538	0,349089027	0,5
SVM (SMO) - RBFKernel - 10^{-5}	69,81780538	0,349089027	0,5
SVM (SMO) - RBFKernel - 10^{1}	81,83022774	0,780230089	0,789593224
SVM (SMO) - RBFKernel - 10^{2}	79,25672878	0,759854624	0,794236236
SVM (SMO) - RBFKernel - 10^{3}	77,68115942	0,745127334	0,780510056
SVM (SMO) - RBFKernel - 10^{4}	77,68115942	0,745127334	0,780510056
SVM (SMO) - RBFKernel - 10^{5}	77,68115942	0,745127334	0,780510056

Tabela 3. Resultados experimentais para Acurácia, Precisão e Recall.

Algoritmo	Acurácia(%)	Precisão	Recall
KNN Coseno K=1	70,67908903	0,711779854	0,751412676
KNN Euclidiana K=1	60,94616977	0,691055755	0,702176147
KNN Coseno K=3	72,83022774	0,736662451	0,781173077
KNN Euclidiana K=3	58,08902692	0,670929719	0,676139856
KNN Coseno K=5	70,39958592	0,716693163	0,75773532
KNN Euclidiana K=5	50,64596273	0,625165438	0,616083533
KNN Coseno K=7	67,8136646	0,694898423	0,730600023
KNN Euclidiana K=7	47,35403727	0,612822203	0,594455954
KNN Coseno K=9	66,52795031	0,683925359	0,71818679
KNN Euclidiana K=9	44,7805383	0,599007306	0,576093491
KNN Coseno K=11	66,24223602	0,685182155	0,718739092
KNN Euclidiana K=11	42,92132505	0,590071367	0,563840098
KNN Coseno K=13	66,38716356	0,684412455	0,718687878
KNN Euclidiana K=13	40,77225673	0,589435145	0,55375898
KNN Coseno K=15	65,8136646	0,679488578	0,712522333
KNN Euclidiana K=15	39,91304348	0,586985449	0,548655333
KNN Coseno K=17	66,09937888	0,683747381	0,717220682
KNN Euclidiana K=17	39,33954451	0,589742623	0,547702999
KNN Coseno K=19	65,52795031	0,678776798	0,711710212
KNN Euclidiana K=19	39,19668737	0,594595802	0,548198215
KNN Coseno K=25	66,09730849	0,67774551	0,710848944
KNN Euclidiana K=25	33,47619048	0,5344712	0,510342164
KNN Coseno K=35	68,66873706	0,682913646	0,71639107
KNN Euclidiana K=35	33,32712215	0,609392508	0,517553368
KNN Coseno K=45	70,81987578	0,688543667	0,721236522
KNN Euclidiana K=45	33,18426501	0,610173465	0,517451077
KNN Coseno K=55	72,67287785	0,699294268	0,728964114
KNN Euclidiana K=55	33,4699793	0,611586966	0,519548979
$IMBHN^C$ taxa 0,01	77,11387164	0,737402008	0,771383853
$IMBHN^C$ taxa 0,05	77,83229814	0,748510131	0,782372099
$IMBHN^C$ taxa 0,1	77,83229814	0,748510131	0,782372099
$IMBHN^C$ taxa 0,5	78,26293996	0,752187208	0,784905539
$IMBHN^R$ taxa 0,01	76,25672878	0,739260395	0,778744039
$IMBHN^R$ taxa 0,05	76,25672878	0,739260395	0,778744039
$IMBHN^R$ taxa 0,1	76,25672878	0,739260395	0,778744039
$IMBHN^R$ taxa 0,5	76,25672878	0,739260395	0,778744039