Análise dos *Tweets* sobre a Universidade do Sul de Santa Catarina – Unisul, por meio da Mineração de Texto e Análise de Sentimentos

Arthur Weber Carlos, Mariana da Rosa Scandolara

Curso de Ciência da Computação - Unidade Acadêmica Ciência da Produção, Construção e Agroindústria – Campus Tubarão – Universidade do Sul de Santa Catarina – Tubarão – SC – Brasil

arthur2weber@gmail.com, marianascandolara@gmail.com

Abstract. *Acknowledging the competition among higher education institutions, the universities are in search of innovating strategies for their activities. This article represents the creation of a computer system that, utilizing text data mining and sentiment analysis, identifies the polarity of opinions sent by Twitter users about the Universidade do Sul de Santa Catarina - UNISUL. The classification is made using the Naive Bayes probabilistic classification. This algorithm uses machine supervised learning techniques and, for the training algorithm, a number of collected tweets were manually classified by a psychologist. Through result validation, it was found a 75,3% efficiency, which can be considered a good performance bearing in mind that what was developed is a prototype.*

Resumo. *Tendo em vista a competitividade entre as instituições de ensino superior, as universidades estão em busca de estratégias inovadoras para suas atividades. Esse artigo apresenta a criação de um sistema computacional que, utilizando a aplicação da mineração de texto e análise de sentimentos, identifica a polaridade das opiniões emitidas pelos usuários do Twitter sobre a Universidade do Sul de Santa Catarina - UNISUL. A classificação é realizada utilizando o classificador probabilístico Naive Bayes. Este algoritmo utiliza técnicas de aprendizagem de máquina supervisionado e, para o treinamento do algoritmo, uma parte dos tweets coletados foram classificados manualmente por uma psicóloga. Através da validação de resultados, foi obtido 75,3% de eficiência, o que pode ser considerado um bom desempenho considerando que foi desenvolvido apenas um protótipo.*

1. Introdução

Com a popularização das redes sociais as pessoas começaram a utilizar essas redes para diversos fins, como se relacionar com outros usuários, se informar e também para emitir suas opiniões sobre diversos assuntos. A inclusão digital permitiu, de acordo com a 11º pesquisa TIC Domicílios 2015, que, no Brasil, 58% da população possuísse acesso à internet (BOCCHINI, 2016).

Hoje, qualquer empresa ou pessoa pode ter interesse em saber o que falam de si na internet. Esse tipo de informação pode servir para tomar decisões assertivas e, assim, gerar mais valor para a sua marca ou nome. Contudo, para transformar esses dados em informações relevantes e identificar se a opinião é positiva ou negativa, encontram-se

algumas dificuldades, como esses conteúdos estarem em forma de texto, o que é conhecido como dados não estruturados, a velocidade e o volume em que novas opiniões são geradas na internet e a capacidade de identificar a polaridade de cada uma. Para sanar essas dificuldades, são utilizadas as técnicas de mineração de texto e análise de sentimentos.

Conforme Mainardes, Ferreira e Tontini (2009) as universidades estão buscando estratégias inovadoras para garantir vantagens comerciais, tendo em vista o crescimento da competitividade no mercado de educação superior. Para basear essas estratégias, são necessários dados sobre a instituição, tradicionalmente obtidos em formato de questionários e análise manual de documentos.

Uma forma inovadora de obter dados opinativos é utilizando as redes sociais. Entre as fontes geradoras de informação mais importantes da atualidade está o *Twitter*. Esta rede social é usada principalmente por pessoas comuns falando sobre assuntos diversos e expressando suas opiniões e experiências.

Com o objetivo de obter informações mais abrangentes, para auxiliar Universidade do Sul de Santa Catarina (UNISUL) em seus processos de tomada de decisões, foi desenvolvido um sistema computacional capaz de classificar em positivo e negativo os *tweets* publicados, através mineração de dados e análise de sentimentos.

O artigo está estruturado da seguinte forma: a segunda seção apresenta a contextualização teórica sobre mineração de texto e análise de sentimentos. A terceira seção refere-se aos trabalhos existentes similares ao sistema desenvolvido. A quarta seção descreve as ferramentas utilizadas no desenvolvimento da solução e o método computacional. A quinta seção apresenta os resultados obtidos com o sistema desenvolvido, bem como as discussões relacionadas a estes resultados. Por fim, a sexta seção apresenta as conclusões relacionadas ao desenvolvimento do sistema e ideias para trabalhos futuros.

2. Contextualização

2.1 Mineração de Texto

A mineração de texto refere-se ao "[...] processo de extração de informações de interesse e padrões não-triviais ou descoberta de conhecimento em documentos de texto não-estruturados" (ARANHA; PASSOS, 2006). Além disso, conforme disciplina Morais (2007), ela traz contribuições na área de busca de informações em documentos e análise qualitativa e quantitativa de grandes volumes de texto, bem como permite uma melhor compreensão de textos disponíveis em variados tipos de documentos.

Moura (2004) considera que o objetivo da mineração de textos é a busca por padrões, tendências e regularidades em textos escritos em linguagem natural.

2.2 Análise de Sentimentos

A análise de sentimentos é uma área que está em constante evolução, devido a seu grande apelo econômico, visto que uma empresa que consegue analisar os comentários em redes

sociais e descobrir a opinião de usuários sobre o seu produto pode focar seus investimentos de uma maneira mais assertiva.

A análise de sentimentos, também chamada de mineração de opinião, pode ser conceituada como o "[...] estudo computacional de opiniões, sentimentos e comoções expressas em texto". (LIU, 2010).

Conforme ensina Ferreira (2010), um texto com opinião possui quatro elementos: o "objeto", para se referir a quem ou ao o que a opinião se destina; o "titular", utilizado para indicar quem expressa a opinião; a "opinião", que é o conteúdo do texto e deste é possível identificar a orientação da opinião, que poderá ser positiva, negativa ou neutra; e o "tempo", que é a data em que a opinião foi expressa.

Conforme Medhat, Hassan e Korashy (2014), as abordagens de análise de sentimento regularmente são divididas em três abordagens: Abordagem de Aprendizagem de Máquina, Abordagem Léxica e Híbrida. Na primeira são utilizados algoritmos de aprendizagem de máquina que são divididos em aprendizagem supervisionado e não supervisionado. Na segunda abordagem a análise é realizada utilizando termos previamente conhecidos e processados. Esta é dividida em abordagem que utiliza dicionário e métodos estatísticos ou semânticos. A terceira abordagem utiliza técnicas de ambas abordagens já mencionadas.

A aprendizagem de máquina pode ser definida, conforme afirma Murphy *apud* Santos (2016), como métodos que identificam automaticamente padrões de relações entre os dados e utilizam como regras gerais para previsões e tomadas de decisões sob incertezas.

2.3 Algoritmo Naive Bayes

O classificador probabilístico Naive Bayes foi escolhido para esse trabalho, que utiliza técnica de aprendizagem supervisionada, por este ser "[...] frequentemente utilizado como base na classificação de textos por ser rápido e fácil de implementar" (RENNIE apud CARVALHO FILHO, 2015).

Teoricamente o "[...] Teorema de Bayes mostra a relação entre uma probabilidade condicional e a sua inversa; por exemplo a probabilidade de uma hipótese dada a observação de uma evidência e a probabilidade da evidência da hipótese" (REISSWITZ, 2009). Considerando que B representa um fato que ocorreu previamente e A um fato que depende de B o algoritmo deve contar o número de vezes que A e B ocorreram juntos e dividir pela quantidade de vezes que B ocorreu sozinho.

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,P(A)}{\mathrm{P}(B)}$$

Equação 1. Teorema de Bayes

A equação 1 ilustra o teorema de Bayes onde P(B|A) é a probabilidade de B dado A, P(A) é a probabilidade a priori de A, e P(B) é a probabilidade a priori de B. Considerando que P(B) deve ser sempre maior que zero.

Contudo, como as palavras se repetiam poucas vezes entre os textos coletados e o universo de palavras variadas era maior, de forma que a divisão ficaria muito pequena, e

alguns casos a palavra testada não constava dentre as palavras avaliadas, foi utilizada uma normalização da probabilidade a posteriori (SUN, 2009).

$$log\frac{P(P|Tw)}{P(N|Tw)} = log\frac{P(P)}{P(N)} + \sum_{i=1}^{n} log\frac{P(p_i+1|P)}{P(p_i+1|N)}$$

Equação 2: Normalização da probabilidade a posteriori

A equação 2 ilustra a Normalização da probabilidade a posteriori, onde P (P|Tw) é a probabilidade a posteriori de um *tweet* pertencer a classe de positivos, P (N|Tw) é a probabilidade a posteriori de um *tweet* pertencer a classe de negativos, P (P) é a probabilidade a priori de um *tweet* ser positivo, P (N) é a probabilidade a priori de um *tweet* ser negativo, n é o número de palavras que compõe o *tweet*, P(pi+1|P) é a probabilidade de cada palavra do *tweet* ser positiva acrescido de um e P(pi+1|N) é a probabilidade de cada palavra do *tweet* ser negativa acrescido de um.

O acréscimo de mais um foi realizado utilizando a técnica de *smoothing* que, de acordo com Gasperin e Lima (2001), "tem como efeito reservar uma pequena porção do espaço de probabilidade para os eventos não conhecidos. ".

3. Trabalhos correlatos

Carvalho Filho (2014) desenvolveu um sistema para analisar os sentimentos que foram expressados no *Twitter* sobre a copa do mundo de 2014. Os *tweet*s foram coletados utilizando um *script* desenvolvido na linguagem *python* onde o autor filtrou *tweet*s que continham termos relacionados à copa do mundo durante o período entre 12/06/2014 a 13/07/2014. Após a coleta dos *tweet*s, foi realizado o pré-processamento. Nesta fase foram removidos os caracteres não alfabéticos, links e nomes de usuários. Também foram removidos as *stopwords* utilizando a plataforma NLTK (*Natural Language ToolKit*). Em seguida foi realizada a mineração de texto, sendo que nesta etapa foi utilizado o Classificador Naive Bayes do Apache Mahout. Com essa pesquisa foi identificado que o classificador apresentou bom resultado para classificar *tweet*s positivos, negativos e neutros. Além disso, apresentou resultado bom para *tweet*s ambíguos.

Nascimento, Osiek e Xexéo (2015) desenvolveram um trabalho de análise de sentimentos em *tweet*s relacionados a notícias. Para realizar esse projeto, os autores coletaram os t*weet*s entre os meses de agosto a outubro de 2011. Neste período, os autores analisaram as notícias no jornal online G1 e relacionaram aos termos presentes nos *trending topics* do *Twitter*. Com essa análise, selecionaram notícias das seguintes categorias: Entretenimento, Política e Policial. De cada categoria foram coletados entre 2 a 3 notícias e para cada uma cerca de 400 *tweet*s. Os *tweet*s coletados foram classificados manualmente pelos autores do artigo. O sistema desenvolvido nesse trabalho utilizou a biblioteca de processamento de texto *LingPipe* e implementou três classificadores: Trigrama; Hexagrama e Naive Bayes. Essa implementação visou identificar qual o melhor classificador. Os autores identificaram que os classificadores testados apresentaram resultados muito similares.

Alves *et al* (2015) desenvolveram um sistema de análise de sentimentos e aplicaram aos *tweet*s relacionados ao meio ambiente. A primeira etapa para o desenvolvimento do trabalho foi a coleta de dados. Foram coletados *tweet*s que continham termos previamente selecionados: Cidade de atuação do Plano de Ação Socioambiental

(PAS); termos relacionados ao meio ambiente e termos relacionados a empresa CHESF. A coleta foi realizada utilizando API disponibilizada pelo *Twitter*, sendo buscados e armazenados os *tweet*s gerados no mês de março de 2015. Após a coleta, foi realizado o pré-processamento. Nesta fase foram removidas as *stopwords*, bem como foram removidos termos comuns no *Twitter* (RT, via), nomes de usuários e tratamento de *hashtags*. Esse trabalho utilizou a abordagem de aprendizado de máquina supervisionado e, para utilizar essa técnica, é necessário criar um corpus de treinamento. Para realizar essa classificação, foi utilizado o classificador probabilístico Naive Bayes. Os resultados obtidos foram considerados satisfatórios, visto que a maioria dos *tweet*s presentes no corpus de treinamento eram objetivos e, portanto, a polaridade neutra. Para a apresentação dos resultados, os autores aplicaram o classificador a todos os *tweet*s coletados e analisaram a polaridade em relação ao tempo.

Santos (2016) desenvolveu um sistema para analisar a polaridade dos *tweet*s publicados sobre a BlackFriday de 2015. Para realizar o experimento, o autor buscou os *tweet*s que continham a palavra BlackFriday em um intervalo de 10 minutos do dia 27 de novembro de 2015. Para a coleta, o autor utilizou a API Tweepy 3.5 do Python, que busca os *tweet*s em *streaming*. Em seguida foi realizado o pré-processamento e as seguintes técnicas foram utilizadas: Remoção dos *Tweet*s Repetidos, Transformação de *Tweet*s em Formato Unicode, Transformação de *Tweet*s para Letras Minúsculas, Transformação de Acrônimos e Abreviações de Internet, Tokenização, Remoção de StopWords e Pontuação e Caracteres Especiais.

A etapa seguinte tratou da mineração de texto. O autor utilizou o algoritmo Naive Bayes implementado pela API NLTK 3.0.4. Para a análise de resultados, o autor verificou que, entre todos os *tweet*s classificados, 51% dos *tweet*s foram classificados como positivos e 49% negativos. O autor também comparou os resultados obtidos com outros trabalhos que avaliaram a opinião de usuários de redes sociais sobre a Black Friday.

4. Materiais e Métodos

4.1 Ferramentas

Para realizar o desenvolvimento do software, foi utilizada como linguagem de programação o PHP 5.2, com a interface sendo programada utilizando CSS 3, HTML 5 e Jquery. Para o banco de dados, foi utilizado o banco de dados relacional PostgreSQL 9.3.

Utilizou-se a ferramenta de gerenciamento de tarefas do Windows para criar uma tarefa agendada que realiza os downloads de *tweet*s e realiza o treinamento e a avaliação de cada *tweet* após ser criado o *trainig corpus*.

4.2 Método computacional

O método computacional utilizado é composto pela mineração de texto e analise de sentimentos. As etapas foram realizadas em sequência conforme ilustrado na figura 1.

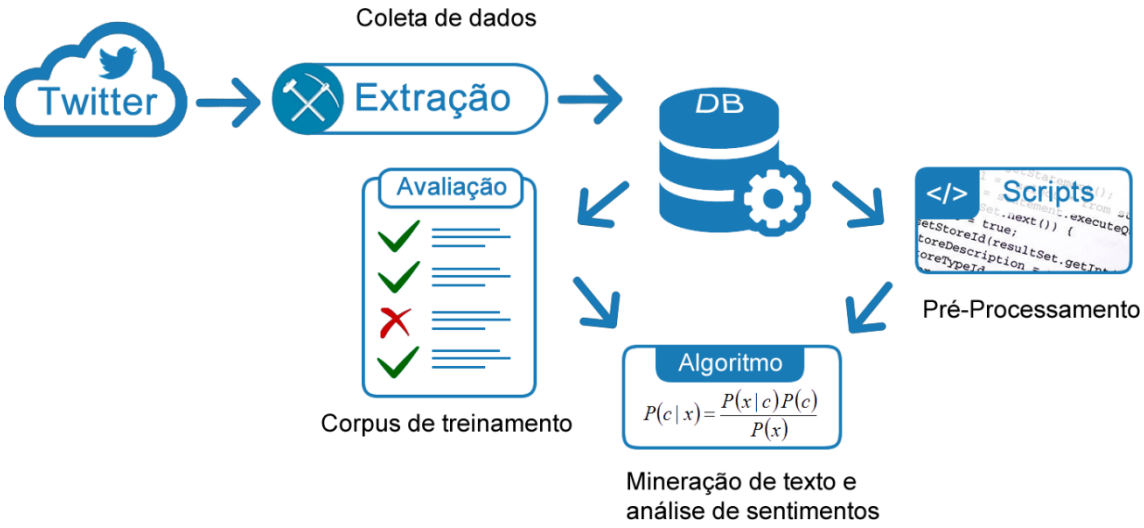

Figura 1. Processo de mineração de texto

4.2.1 Mineração de texto

A mineração de texto é realizada em etapas, que comumente são divididas em quatro: coleta, pré-processamento, indexação, e análise de informações (ARANHA; 2007).

Coleta

A primeira etapa é a coleta. Conforme aduz Aranha (2007), ela tem como objetivo coletar a base textual que será utilizada para a mineração de texto.

Neste projeto, para criação do sistema, na primeira etapa, realizou-se a busca dos dados do *Twitter*. Apesar de mineração de texto utilizar normalmente bancos de dados não-relacional por minerar dados que não possuem uma estrutura definida, neste trabalho foi definido utilizar um banco de dados relacional, devido a fonte de dados possuir como característica a regra que restringe os *tweet*s a no máximo 140 caracteres.

O *Twitter* disponibiliza uma API para a busca de *tweets* em sua base, porém a API é restrita a buscar dados de até uma semana atrás. Como esse projeto tem por objetivo buscar dados históricos da Unisul no *Twitter,* criou-se um *script* para buscar os *tweet*s que contenham uma palavra específica, em qualquer período de tempo selecionado pelo usuário desde a criação da rede social 15/06/2006. Desses *tweet*s, foram recuperadas as seguintes informações: o texto do *tweet*, a identificação de quem postou e o dia que foi postado.

Coletou-se 60.113 *tweet*s que continham a palavra UNISUL e estes foram postados entre o dia 15/06/2006 e o dia 16/04/2017. Após a coleta, realizou-se o pré-processamento que eliminou alguns *tweet*s por estarem fora de contexto. Foram eliminados 18.850 e permaneceram 41.263 *tweet*s disponíveis na base para avaliar os sentimentos dos usuários do *Twitter* sobre a UNISUL.

Pré-Processamento

A segunda etapa é o pré-processamento. De acordo com Carrilho (2007), ele tem como objetivo prover alguma formatação sobre a massa textual coletada.

Na primeira etapa do pré-processamento, selecionou-se os *emoticons* mais comuns, por exemplo ":)", "<3", identificou-se em quais *tweet*s os mesmos estavam presentes e coletou-se os mesmos para uma coluna de *emoticons* de cada *tweet*. Em seguida, removeu-se os caracteres não alfanuméricos, pois esses símbolos não devem ter polaridade a menos que façam parte de um *emoticon*.

Também foram aplicadas as seguintes técnicas de Processamento de Linguagem Natural (PLN): Normalização de palavras, esta pode ser realizada aplicando diversas métodos, nesse trabalho foram aplicados: Acentuação, onde foram removidas as acentuações das palavras e Letras Maiúsculas onde todas as letras são transformadas em minúsculas, para que não haja a diferenciação pelo software das palavras começadas com letra maiúscula (GOMES, 2012).

Após a normalização do texto foi realizada a Tokenização, a fim de segmentar os termos contidos no texto e "traduzir um texto em dimensões possíveis de analisar" (GOMES, 2012). Exemplo disso seria a frase "A UNISUL é a melhor universidade do país" tokenizada para os termos {"A", "UNISUL", " é", "a", "melhor", "universidade", "do", "país"}.

Em seguida, realizou-se a Remoção de Palavras não Discriminantes (*stop words*), que tem por objetivo remover palavras que não agregam significado ao texto, como, por exemplo, preposições, conjunções, artigos, etc.. A remoção dessas palavras visa diminuir o tempo de processamento (GOMES, 2012). Por exemplo, "A UNISUL é a melhor universidade do país" sem as stop words, reduz-se a "UNISUL melhor universidade país".

Além disso, analisou-se manualmente uma amostra de *tweet*s e identificou-se alguns assuntos continham a palavra UNISUL e não tinham relação com a universidade. Após identificados esses assuntos foram removidos da base de dados por scripts. Os assuntos identificados foram futsal, que se referiria ao time de futebol da UNISUL, UNISUL TV, que se refere ao canal de televisão da UNISUL, *tweets* gerados pelas contas da própria universidade, que continham marketing e por isso não pode ser considerado a opinião de um usuário sobre a instituição, e também *tweet*s de *check in* e *tweet*s de aviso de postagem em outras redes sociais.

Indexação

Conforme Aranha (2007), a indexação é a etapa que tem por objetivo indexar os dados e facilitar a busca de informações no texto.

Neste projeto essa etapa não foi aplicada, pois esse projeto não tem por objetivo a recuperação de informação e sim a análise de sentimentos.

Corpo de treinamento

Após o pré-processamento, criou-se o *training corpus* (corpo de treinamento), que foi uma seleção aleatória de *tweet*s. Estes foram disponibilizados, pelo sistema desenvolvido, conforme mostrado na Figura 2, para um profissional da área de psicologia, para que o mesmo classificasse em positivo e negativo. Foram avaliados 280 *tweet*s e estes foram classificados em 203 *tweet*s como positivos e 77 como negativos.

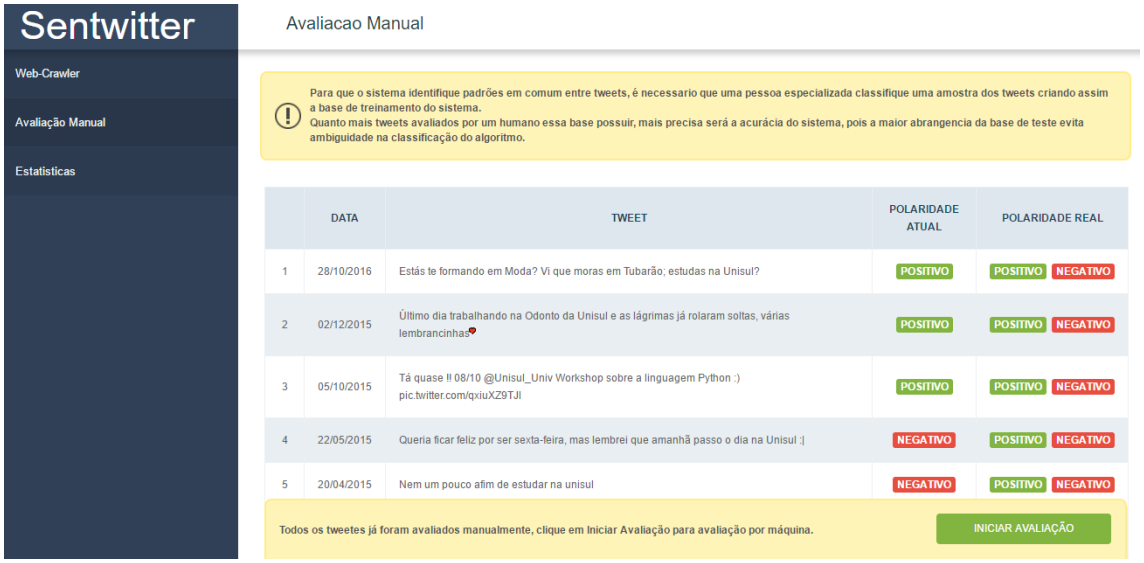

Figura 2. Tela de avaliação manual

Procedimento de testes

Como mencionado, este trabalho utilizou o classificador probabilístico Naive Bayes. Nesta etapa o algoritmo foi executado e, para identificar a acurácia do teste, realizou-se o procedimento de validação cruzada. Nele, os *tweet*s do *traing corpus* foram subdivididos em 10 grupos de 28 *tweet*s cada um. Nove grupos foram utilizados para treinar o algoritmo e um foi utilizado para testá-lo. Esse procedimento foi repetido dez vezes para que todos os grupos fossem utilizados para teste no mínimo uma vez. Somou-se o resultado dos testes, que foram aplicados a uma matriz de confusão para que os resultados pudessem ser analisados de uma forma clara.

Após o treinamento, aplicou-se o algoritmo para todos os *tweet*s coletados, e o mesmo identificou a polaridade em cada um.

Analise de resultados

Por fim, a última etapa da mineração de textos é a análise de resultados. De acordo com Aranha (2007), os resultados obtidos podem ser medidos, por exemplo, pelo tempo de processamento, pela compatibilidade dos resultados de acordo com um especialista da área e pela devida aplicabilidade dos resultados recuperados.

Neste projeto, utilizou-se como métrica para análise dos dados a matriz de confusão, um tipo de tabela que permite a visualização do desempenho de um algoritmo supervisionado. Com a sua utilização se torna mais simples verificar se o algoritmo está confundindo as classificações. As classificações corretas ficam localizadas na diagonal principal e os palpites errados nas outras células da tabela (REVISTABW, 2015). Utilizando essa métrica foi possível identificar a número de acertos em classificar os *tweet*s nas classes positivo e negativo. Também foi possível mediar as métricas de acurácia, sensibilidade, especificidade, eficiência, valor preditivo positivo, valor preditivo negativo, e o coeficiente de correlação de Matthews.

O Coeficiente de correlação de Matthews (PHI) é considerado uma medida balanceada que pode ser usada mesmo quando as classes de estudo (*tweet*s positivos e

*tweet*s negativos) são de tamanhos muito desiguais. O coeficiente assume valores entre [-1 e +1], onde um valor igual a +1 corresponde a predição prefeita, 0 corresponde a predição completamente aleatória e -1 a predição inversa

5. Resultados e discussões

5.1 Scores da Análise de Sentimento

A tabela 1 apresenta uma amostra de *tweets* que foram avaliados pelo modelo proposto, o treinamento do modelo foi realizado com o training corpus avaliado previamente.
A coluna Tweet inclui o texto original do *tweet*, a coluna Score informa o valor obtido na aplicação do teorema de Bayes com a normalização da probabilidade a posteriori, e a coluna Orientação Obtida contém a classificação que o modelo gerou. O modelo foi programado para considerar um *tweet* negativo quando o score for menor que 0 (zero), e considera-lo positivo quando o score for igual ou maior que 0 (zero).

Tabela 1. Amostra de *tweets*, score e orientação

Tweet	Score	Orientação Obtida
Ganhei altos copinho da Unisul	0.13366359556660150285	Positivo
tanta coisa p esquecer na unisul e eu esqueço logo o carregador do celular	-0.03970892884420302080	Negativo
fiquei 6 meses tentando entender o site da unisul, qnd finalmente consigo... eles mudam	-0.07601154681461153407	Negativo
Nem um pouco afim de estudar na unisul	-0.000000009965943099344830298	Negativo
Sdds de joga na unisul :/	-0.06420202303097332015	Negativo
muito legal hoje, fomos pra pista e depois pra unisul	0.15165796934128349525	Positivo
Massa amanhã vou ter q ficar das 7 até as 22:30 na unisul e to sem um real no bolso e mãe braba cmg no deixou nd de dinheiro em casa ☞🤦‍♂️‍♀️😩	-0.22085864443511232213	Negativo
Mais duas semanas de férias na Unisul, por incrível que pareça estou sentindo falta! Só quero ver ano que vem.. hahahaha	0.57692475841655261178	Positivo
Unisul, praça de alimentação. Quero minha casaaaa, cuidar da minha mããããe. :~	0.12296324344585575725	Positivo
ta decidido e vou ir p Unisul mesmo	0.00910800445287427304	Positivo

5.2 Análise dos Resultados

Na matriz de confusão demonstrada na tabela 2, a diagonal principal representa os acertos da rede, sendo 159 acertos para positivo e 42 acertos para negativos. A diagonal inversa representa os erros preditos pelo classificador, sendo que em 35 casos o classificador inferiu positivo para *tweet*s negativos (falso-positivo) e em 44 casos a rede inferiu negativo para *tweet*s positivos (falso-negativo).

Tabela 2. Matriz de confusão experimento 1

	Classificação manual		
Classificação automática		Positivo	Negativo
	Positivo	159	35
	Negativo	44	42

Para analisar os dados obtidos neste experimento, utilizaram-se algumas métricas que visam identificar a qualidade das informações.

A acurácia é uma medida que calcula a proporção de predições corretas e é calculada pela equação 3.

$$ACURÁCIA = \frac{VP + VN}{(P + N)}$$

Equação 3. Acurácia

VP é o número de acertos positivos, VN é o número de acertos negativos, P e N é o número de *tweet*s classificados como positivos e negativos, respectivamente, pela avaliação manual. Neste experimento, atingiu-se a acurácia de 78,3%.

A sensibilidade é uma métrica que mede a proporção de *tweet*s positivos serem classificados pelo algoritmo como positivo. Neste experimento a sensibilidade foi de 78,3 %. A sensibilidade é calculada com a equação 4.

$$SENSIBILIDADE = \frac{VP}{P}$$

Equação 4. Sensibilidade

Outra métrica utilizada foi a especificidade que mede a proporção de *tweet*s negativos serem classificados corretamente como negativos. Neste experimento, a especificidade foi de 54,5%, e é calculada conforme a equação 5.

$$\text{ESPECIFICIDADE} = \frac{VN}{N}$$

Equação 5. Especificidade

A métrica da eficiência é a média aritmética da sensibilidade e especificidade. Quando o conjunto de dados se encontra desbalanceado é comum gerar muitos casos de falso-positivo e falso-negativo. Essa métrica é utilizada a fim de balancear os resultados obtidos, e neste experimento obteve índice de 66,4%. Ela é obtida através da equação 6.

$$\text{EFICIÊNCIA} = \frac{(\text{SENSIBILIDADE} + \text{ESPECIFICIDADE})}{2}$$

Equação 6. Eficiência

Valor Preditivo Positivo (VPP) mede a proporção de *tweet*s positivos, dado que o algoritmo assim os identificou. Neste experimento foi obtido 82,0% e é calculado pela equação 7.

$$VPP = \frac{VP}{(VP + FN)}$$

Equação 7. Valor Preditivo Positivo

Valor Preditivo Negativo (VPN) também mede a proporção de *tweet*s, porém dos negativos, dado que o algoritmo assim os identificou. Neste experimento foi obtido 48,8% e é calculado pela equação 8.

$$VPN = \frac{VN}{(VN + FP)}$$

Equação 8. Valor Preditivo Negativo

Como essas medidas são suscetíveis a induzir a conclusão errada sobre o desempenho do sistema dependendo do desbalanceamento do conjunto de dados, foi utilizada a métrica que mede o coeficiente de correlação de Matthews para ter uma visão da qualidade dos dados.

Neste experimento o Coeficiente de correlação de Matthews, obteve-se o valor de 0,32. Este coeficiente é dado pela equação 9.

$$phi = \frac{(VP * VN - FP * FN)}{\sqrt{(VP + FP) * (VP + FN) * (VN + FP) * (VN + FN)}}$$

Equação 9. Coeficiente de Correlação de Matthews

Como o resultado obtido pelo coeficiente de Matthews foi abaixo de 0,50, realizou-se outro teste para diminuir o desbalanceamento do conjunto de dados. Dentro dos *tweet*s do *training corpus,* foram selecionados 77 *tweet*s classificados como negativos, que eram todos que continham essa classificação, e 77 *tweet*s classificados como positivos, sendo que estes foram selecionados aleatoriamente entre os *tweet*s que continham essa classificação.

Com esses 154 *tweet*s, realizou-se uma nova validação cruzada. Porém, nessa classificação, os *tweet*s foram divididos em 11 grupos de 14 *tweet*s cada. O procedimento de treinamento e teste foi feito 11 vezes para que todos os grupos fossem utilizados uma vez para teste. Após a validação, somou-se os valores obtidos e aplicados a uma matriz de confusão, apresentado na tabela 3, para analisar os dados obtidos.

Tabela 3. Matriz de confusão experimento 2

	Classificação manual		
Classificação automática		Positivo	Negativo
	Positivo	49	10
	Negativo	28	67

A acurácia obtida neste experimento foi de 75,3%, que é 3,5% maior que a acurácia alcançada do experimento com os dados desbalanceados. A sensibilidade foi de 63,6%, que é 14,7% menor que a sensibilidade obtida no experimento anterior. Isto se deu por aumentar a proporção de *tweet*s falso-positivo neste experimento. A especificidade foi de 87,0%, um aumento de 32,5%. Este índice foi obtido em razão do número de falsos-positivos ter diminuído neste experimento.

O VPP obtido neste experimento foi de 83,1%, apenas 1,1 % maior que o valor obtido no experimento anterior. Ademais, obteve-se no índice de VPN o valor de 70,5%, um aumento de 51,7% comparado ao experimento anterior. A eficiência do segundo experimento foi de 75,3 %, que é 8,9% maior que a eficiência obtida no experimento anterior.

Todos esses índices podem induzir a conclusões erradas sobre o desempenho do classificador quando os dados de entrada estão desbalanceados. Contudo, o segundo experimento foi executado com o mesmo número de dados positivos e negativos e, por isso, seus índices representam o desempenho do classificador.

Para que seja possível comparar o desempenho de ambos experimentos utilizando um índice balanceado, utilizou-se o coeficiente de Matthews, que no segundo experimento obteve o valor 0,52 conforme exposto na tabela 4.

Tabela 4. Tabela de comparação de índices entre o experimento 1 e o experimento 2

Incide	Experimento 1	Experimento 2
Acurácia	71,8 %	75,3%
Sensibilidade	78,3%	53,6%
Especificidade	54,5%	87,0%
Valor Preditivo Positivo	82,0%	83,1%
Valor Preditivo Negativo	48,8%	70,5%
Eficiência	66,4%	75,3%
Coeficiente de Correlação de Matthews	0,32	0,52

Comparando os dados obtidos, identificou-se que, quando as classes estavam desbalanceadas os *tweet*s foram classificados como positivos com mais frequência, tanto em *tweet*s verdadeiros positivos quanto em falso-positivo. Quando o experimento foi feito com dados balanceados, os *tweet*s tenderam a ser classificados negativamente pelo sistema com mais frequência.

Analisando os dados classificados, identificou-se que em alguns casos uma palavra positiva estava presente em um *tweet* positivo e a mesma palavra estava presente em outro *tweet* que tinha contexto negativo. Por exemplo, a palavra férias que está nos seguintes *tweet*s: "Mais duas semanas de férias na Unisul, por incrível que pareça estou sentindo falta! Só quero ver ano que vem.. hahahaha" e "Sério q eu tenho q ir p unisul mesmo agr nas férias? 😒". Percebe-se que, enquanto o primeiro tem a sua polaridade positiva, o segundo tem a polaridade negativa.

Como o número de palavras é muito grande e se repetem poucas vezes dentro dos *tweet*s avaliados, algumas palavras tiveram seu sentido alterado. Essa deficiência dos experimentos pode ser amenizada aumentando o volume de dados de treinamento.

6. Conclusões

Esse trabalho apresentou um modelo de mineração de texto e análise de sentimentos e os utilizou para minerar a opinião de usuários sobre a Universidade do Sul de Santa Catarina – UNISUL. Utilizando-se a análise de sentimentos, os *tweet*s coletados foram classificadas em positivo ou negativo.

O modelo obteve uma considerável taxa de eficiência em um treinamento com dados balanceados, o que é bom para um protótipo de classificador que não considera o contexto em que as postagens foram feitas, e possui um *corpus* de treinamento relativamente pequeno se comparado ao número de palavras da língua portuguesa.

O sistema computacional desenvolvido possui uma tela de avaliação manual, para que nesta seja possível treinar o algoritmo por uma pessoa de uma área leiga em computação. Neste experimento, houve a participação de uma psicóloga para realizar o treinamento do *training corpus*.

Apesar de ter sido buscada apenas a palavra "UNISUL" dentro do período de tempo de 15/06/2006 16/04/2017 para realizar o experimento, o sistema desenvolvido permite buscar outras palavras chave e em outros períodos de tempo.

Esse projeto trouxe como inovação a busca de dados históricos dentro do *Twitter* buscando todos os *tweet*s já gerados que continham a palavra buscada e não apenas os *tweet*s disponibilizados pela API do *Twitter*. Isto pode ser bastante útil para empresas que já tenham realizado alguma campanha de marketing no passado e queira identificar a opinião dos usuários do *Twitter* na época.

Para implementar essa solução em outra empresa seria necessário realizar um trabalho de consultoria, para que possa ser analisado os termos utilizados nos *tweet*s coletados e restringir os assuntos coletados aos assuntos de interesse da marca pesquisada.

Para trabalhos futuros se pretende criar uma funcionalidade que permita alimentar a base de dados em *streaming* após a coleta inicial de dados históricos e, na etapa de visualização das estatísticas, mostrar os textos dos *tweet*s para que seja possível identificar qual característica da universidade especificamente as pessoas estão se referindo ao emitir as opiniões.

Referências

ALVES, André Luiz Firmino et al. . Uso de Técnicas de Análise de Sentimentos em *Tweet*s relacionados ao Meio-Ambiente. In: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS - WCAMA, 6. 2015, Recife. Anais do XXXV Congresso da Sociedade Brasileira de Computação, 2015.

ARANHA, Christian; PASSOS, Emmanuel. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação,** [S.I.], p.1-8, dez. 2006. Disponível em: <http://www.periodicosibepes.org.br/ojs/index.php/reinfo/article/view/171/66>. Acesso em: 01 set. 2016.

ARANHA, Christian Nunes. Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. 2007. 144 f. Tese (Doutorado) - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

BOCCHINI, Bruno. **Pesquisa mostra que 58% da população brasileira usam a internet.** 2016. Disponível em: <http://agenciabrasil.ebc.com.br/pesquisa-e-inovacao/noticia/2016-09/pesquisa-mostra-que-58-da-populacao-brasileira-usam-internet>. Acesso em: 20 maio 2017.

CARRILHO JUNIOR, João Ribeiro. **Desenvolvimento de uma Metodologia para Mineração de Textos.** 2007. 96 f. Dissertação (Mestrado) - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

CARVALHO FILHO, José Adail. Mineração de textos: análise de sentimentos utilizando *Tweet*s referentes à Copa do Mundo. 2014. 44 f. TCC (Graduação) - Curso de Engenharia de Software, Universidade Federal do Ceará, Quixadá, 2014.

FERREIRA, Emanuel de Barros Albuquerque. **Análise de Sentimento em Redes Sociais Utilizando Influência das Palavras.** 2010. 69 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade Federal de Pernambuco, Recife, 2010.

GASPERIN, Caroline Varaschin; LIMA, Vera Lúcia Strube de. **Fundamentos do Processamento Estatístico da Linguagem Natural.** Porto Alegre: Faculdade de Informática – Pucrs, 2001. 57 p. (N° 021). Disponível em: <http://www.pucrs.br/facin-prov/wp-content/uploads/sites/19/2016/03/tr021.pdf>. Acesso em: 20 maio 2017.

GOMES, Helder Joaquim Carvalheira. Text Mining: Análise de Sentimentos na classificação de notícias. 2012. 60 f. Dissertação (Mestrado) - Curso de Especialização em Gestão do Conhecimento e Business Intelligence, Instituto Superior de Estatística e Gestão de Infor Mação, Universidade Nova de Lisboa, Lisboa, 2012.

LIU, Bing. Sentiment Analysis and Subjectivity. In: INDURKHYA, Nitin; DAMERAU, Fred J. **Handbook of natural language processing.** 2. ed. Nova Iorque: Crc Press, 2010. p. 627-666.

LIU, Bing. **Web Data Mining:** Exploring Hyperlinks, Contents, and Usage Data. 2. ed. Nova Iorque: Springer, 2011. 622 p.

MAINARDES, Emerson Wagner; FERREIRA, João; TONTINO, Gerson. Vantagens Competitivas em Instituições de Ensino Superior: Proposta e Teste de um Modelo. In:

Colóquio Internacional sobre Gestão Universitária na América do Sul,9., 2009. Florianópolis: Disponível em: <https://repositorio.ufsc.br/xmlui/handle/123456789/36803>. Acesso em: 01 set. 2016.

MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, [S.I.], v. 5, p.1093-1113, maio 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2090447914000550>. Acesso em: 01 set. 2016.

MORAIS, Edison Andrade Martins. Contextualização de Documentos em Domínios Representados por Ontologias Utilizando Mineração de Textos. 2007. 110 f. Dissertação (Mestrado) – Instituto de Informática, Universidade Federal de Goiás, Goiânia, 2007.

MOURA, M. F. Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos. Embrapa Informática Agropecuária, 2004, ISSN 1677-9274, 2004.

NASCIMENTO, Paula; OSIEK, Bruno Adam; XEXÉO, Geraldo. ANÁLISE DE SENTIMENT O DE *TWEET*S COM FOCO EM NOTÍCIAS. Revista Eletrônica de Sistemas de Informação, [S.l.], v. 14, n. 2, ago. 2015.

REISSWITZ, Flavia. Análise De Sistemas. Joinville: Clube de Autores, 2009. 43 p.

REVISTABW. Matriz de Confusão. 2015. Disponível em: <http://www.revistabw.com.br/revistabw/matriz-de-confusao/>. Acesso em: 10 jun. 2017

SANTOS, Wilian Pereira da Silva. Análise dos *Tweet*s sobre a Black Friday através da Mineração de Texto e Análise de Sentimentos. 2016. 50 f. TCC (Graduação) - Curso de Sistemas de Informação, Centro de Ciências Exatas e Tecnologia Escola de Informática Aplicada, Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

SUN, Tianhao. (2009). Spam Filtering based on Naive Bayes Classification. Unpublished master's thesis .