Modelos de Classificação de Sentimentos em *Tweets* usando Programação Genética

Airton Bordin Junior[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Caixa Postal 131 – 74690-900 – Goiânia – GO – Brazil

Abstract. *The increase in the number of Internet users in recent years has resulted in a growing content production by its users. Often, the WEB is used as a platform for debates, opinions, evaluations, etc. This fact, in line with the ease of obtaining the information, made the area of Sentiment Analysis, also called Opinion Mining, a growing interest on the part of researchers.*
[continue]

Resumo. *O aumento no número de usuários de Internet nos últimos anos teve como consequência uma crescente produção de conteúdo por seus usuários. Frequentemente, a WEB é utilizada como plataforma para debates, opiniões, avaliações, etc. Esse fato, alinhado a facilidade de obtenção dessas informações, fez com que a área de Análise de Sentimentos, também chamada de Mineração de Opiniões, tivesse um interesse crescente por parte de pesquisadores.*
[continuar]

1. Introdução

A Análise de Sentimentos, também conhecida como Mineração de Opiniões, é uma linha de pesquisa que tem por objetivo a classificação das emoções de um determinado texto, geralmente como positivo, negativo ou neutro. A área vem ganhando destaque nos últimos anos, principalmente por conta da popularização do acesso à Internet e o consequente aumento na quantidade de conteúdo produzido na rede. O uso das redes sociais como *Twitter* e Facebook e a forma com que as pessoas compartilham suas opiniões e sentimentos sobre os mais diversos assuntos tem motivado a pesquisa de formas automatizadas de classificação desses textos.

Podemos dividir as abordagens de classificação de sentimentos em duas classes principais: técnicas supervisionadas e não supervisionadas. A primeira delas utiliza abordagens de aprendizado de máquina para a classificação das opiniões, realizando o treinamento com mensagens previamente classificadas. A abordagem não supervisionada atua em aspectos estruturais do texto e frequentemente fazem uso de Dicionários Léxicos.

Para que um classificador tenha resultados satisfatórios, deve levar em consideração aspectos inerentes do contexto pertencentes às opiniões que serão avaliadas. Um modelo de classificação de *Tweets*, por exemplo, geralmente é diferente de um processo de classificação de avaliações de produtos ou comentários políticos.

A criação de um modelo de classificador de sentimentos de forma manual também depende de conhecimento prévio sobre o domínio a ser analisado, e demanda experiência do projetista, o que pode prejudicar as estratégias e abordagem do sistema.

A programação Genética, explanada em detalhes no seção 2.2, é uma área da computação evolucionária que busca a criação de modelos para resolver problemas de forma automatizada. Um classificador de sentimentos pode ser visto como um modelo, e a criação de um classificador pode ser abordada como um problema de otimização. Com isso, o presente artigo trabalha com a hipótese que a utilização de Programação Genética apresenta-se como uma alternativa viável para o treinamento e a criação de modelos de classificação de sentimentos eficientes e com resultados satisfatórios.

A proposta principal deste trabalho é a criação de um sistema para a geração de um modelo de classificação de sentimentos, aderente ao contexto para o qual foi treinado, fazendo uso de Programação Genética.

Algumas questões de pesquisa orientam o desenvolvimento do artigo, e a intenção é que ao final do desenvolvimento do mesmo possamos respondê-las. São elas:

- É possível criar, de forma automatizada, um modelo de classificação de sentimentos utilizando Programação Genética?
- Caso seja possível, esse modelo é eficiente na classificação de *Tweets* para o contexto que foi criado?
- Os resultados desse modelo são compatíveis com os classificadores disponíveis atualmente?

A criação de um sistema para a geração automatizada de modelos de classificação de sentimentos, para contextos específicos, e com resultados satisfatórios, pode ser útil para integrar as soluções presentes na literatura e apoiar na validação e na melhoria de modelos já existentes.

Este trabalho está organizado da seguinte maneira: inicialmente, conceitos essenciais para o entendimento do problema de pesquisa são apresentados. Na sequência, a abordagem da solução é apresentada em detalhes. Trabalhos relacionados são discutidos de forma a permitir a comparação do método proposto com outros existentes na literatura. Por fim, os resultados do processo são apresentados.

2. Conceitos

Nesta seção, serão apresentados, de forma sucinta, conceitos fundamentais para o entendimento do trabalho. Pontos principais dos temas serão discutidos, com foco nos conceitos relevantes para a solução do problema de pesquisa.

2.1. Análise de Sentimentos

A Análise de Sentimentos, também chamada de Análise de Opiniões ou Mineração de Opiniões, é uma linha de pesquisa abrangente e que vem sendo tema de diversos trabalhos nos últimos anos. Como observado em [Liu 2010], esse crescente interesse sobre o assunto ocorre principalmente devido ao aumento no número de usuários de Internet e o consequente crescimento da produção de conteúdo independente na rede, como opiniões, avaliações, entre outros.

Essa área de estudo tem como principal desafio a Análise de Opiniões, descritas em linguagem natural, para a identificação da polaridade implícita ou explícita no texto. Essa polaridade é, na maior parte das vezes, identificada como uma escala de pontuação de sua característica positiva, negativa ou neutra.

Quanto aos classificadores de sentimentos, há duas formas principais para o processamento e classificação das mensagens: a abordagem supervisionada e não supervisionada. Na abordagem supervisionada, técnicas de aprendizado de máquina são aplicados à mensagens previamente rotuladas de forma a identificar características que auxiliem na distinção e detecção de sentimentos nas sentenças desconhecidas. Técnicas não supervisionadas atuam principalmente em características sintáticas e semânticas do texto e, geralmente, baseiam-se em Dicionários Léxicos - conjunto de palavras e suas polaridades (grau de positividade e negatividade de uma mensagem). À partir desse dicionário, é feito o processamento das mensagens pelo classificador e retornada a polaridade da mesma. [Araújo et al. 2013]

Dentre as principais técnicas de aprendizado de máquina para a classificação de sentimentos, podemos citar o *Support Vector Machines* (SVM) [Haddi et al. 2013], *Naïve Bayes* [Iqbal et al. 2015], *Adaboost* [Graff et al. 2017], Redes Neurais Artificiais, entre outros. [Rodrigues et al. 2016]

Importante destacar que, para o desenvolvimento de um bom modelo supervisionado, é essencial que haja uma quantidade considerável e representativa de dados previamente rotulados pertencentes ao domínio do problema, para que o treinamento possa ser feito de forma satisfatória. [Araújo et al. 2013]

As abordagens não supervisionadas baseadas em Léxico, como discutido anteriormente, fazem uso de um conjunto de palavras e suas polaridades. Alguns classificadores utilizam Dicionários Léxicos já existentes, enquanto outros se encarregam de criar o próprio dicionário, mais adequado ao contexto e domínio da análise. Essa estratégia faz uso das características do texto e regras sintáticas para determinar a classificação de palavras e frases desconhecidas à partir das palavras contidas no dicionário e suas relações. Dentre as principais técnicas dessa categoria podemos citar *Part-of-Speech Tag* [Becker et al. 2013], *Pointwise Mutual Infomation* (PMI) [Turney 2002], entre outros.

Para técnicas que se encarregam da criação e expansão do próprio Dicionário Léxico há, basicamente, 3 formas de fazê-lo: manualmente - processo realizado por especialistas humanos que analisam cada palavra, atribuindo uma Orientação Semântica para cada uma delas - e duas formas (semi) automatizadas: baseada em Dicionário e baseada em Corpus. Frequentemente, essas técnicas são utilizadas em conjunto, principalmente a validação manual de Dicionários criados de forma automatizada.

O contexto no qual uma palavra está inserida muitas vezes determina seu valor na classificação da opinião como um todo. Palavras pode ter polaridades diferentes, dependendo do domínio ao qual é aplicada. Sabendo disso, classificadores de sentimentos devem levar em consideração uma série de fatores para que os resultados da mineração de opiniões seja satisfatório. Não há um modelo pré definido para a classificação de sentimentos em qualquer contexto.

2.2. Programação Genética

Programação Genética é um campo da computação evolucionária que busca resolver problemas, de forma automatizada, sem demandar conhecimentos detalhados sobre a solução [Koza 1992]. De forma geral, podemos definir a Programação Genética como um método sistemático, não dependente de um domínio específico, usado para permitir que computadores criem programas para solução de problemas de forma automática, inciando com um

conhecimento de alto nível sobre as regras gerais dos possíveis modelos.

Nesse contexto, programa significa um modelo capaz de, à partir de uma ou mais entradas, produzir uma saída para as mesmas. Embora possam ser representadas por diversos tipos de estruturas, a forma mais comum de representação é por meio de árvores, onde os nós internos representam funções e os nós folha representam terminais do problema. Um exemplo de um programa pode ser visto na figura 1, que representa o código *if(X >Y) then { X * 6 + 1.9 } else { X / cos(X) }*

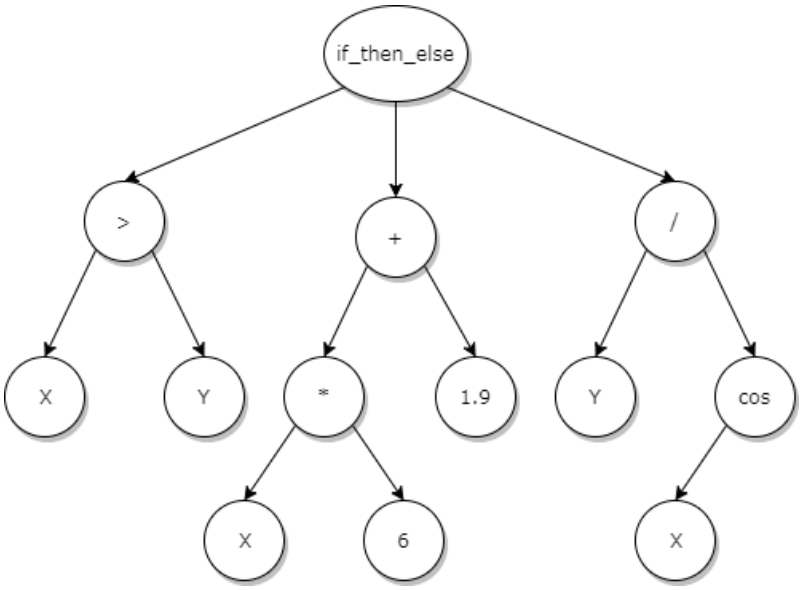

Figure 1. Exemplo de um programa em Programação Genética

Na Programação Genética, assim como em outros algoritmos baseados na evolução humana, são criadas populações onde cada indivíduo representa uma possível solução para o problema. A forma mais comum de inicialização da população é fazê-las de forma aleatória, evoluindo as mesmas no decorrer dos ciclos, chamados de gerações. Para cada geração, programas possivelmente melhores são criados, evoluindo os programas (modelos) gerados. Assim como a natureza, a Programação Genética é um processo aleatório, e não garante o resultado ótimo. Porém, essa aleatoriedade faz com que, muitas vezes, as soluções fujam de problemas como soluções máximas locais, frequentemente enfrentados por métodos determinísticos gulosos [McPhee et al. 2008].

Uma das características mais importantes da Programação Genética é a função de aptidão, ou função *fitness*. Essa função busca representar, de forma quantitativa, a similaridade de cada indivíduo em relação ao resultado esperado. A escolha de uma boa função *fitness* está intimamente ligada ao tipo do problema.

Os responsáveis pela evolução da população de indivíduos são os operadores genéticos. Para a Programação Genética, os principais operadores são a seleção, mutação e *crossover*. Na seleção, um indivíduo é escolhido para fazer parte da próxima geração. A mutação modifica um nó da árvore, de forma a criar um indivíduo modificado em uma das partes escolhida aleatoriamente. O *crossover* realiza o cruzamento entre dois indivíduos (pais), gerando duas novas possíveis soluções para o problema (filhos). Além desses principais operadores, existem outros como a edição, encapsulamento e a destruição

[PATELLI 2011].

Por sua característica inerentemente paralela, esse tipo de abordagem consegue encontrar resultados muito próximos da solução ótima (às vezes encontra a melhor resolução) para problemas complexos com grandes espaços de busca.

3. Trabalhos relacionados

Como discutido em seções anteriores, a quantidade de trabalhos na área de análise de sentimento vem crescendo a cada ano, motivado, principalmente, pela importância da área no contexto atual de geração e análise de grande quantidade de dados e informações.

Ao debatermos os trabalhos relacionados à área de mineração de opiniões, é praticamente impossível não iniciarmos citando [Liu 2010], uma das principais referências do assunto. O autor, um dos principais nomes sobre o assunto, conceitualiza o problema e propõe uma forma estruturada de organização dos dados não estruturados, característica instrínseca dos textos em linguagem natural, objeto de entrada da pesquisa. A definição de opinião como uma quíntupla (entidade, aspecto da entidade, sentimento, autor e tempo) é utilizada em grande parte dos trabalhos na área, caracterizando-se, portanto, como elemento fundamental nas pesquisas sobre o assunto. Visão geral sobre o tema e principais desafios e técnicas são vistos também em [Mohammad 2016], [Ghaleb and Vijendran 2016], [Guimaraes et al. 2016], [Taboada et al. 2011], [Bandhakavi et al. 2016], [D'Andrea et al. 2015], [Kaji and Kitsuregawa 2007], entre outros trabalhos.

Embasamento teórico sobre a divisão dos métodos de classificação de sentimentos em abordagens supervisionadas e não supervisionadas são apresentadas de forma clara em [Araújo et al. 2013]

No contexto de prognóstico automatizado de Orientação Semântica de palavras, um dos primeiros trabalhos apresentados foi [Hatzivassiloglou and McKeown 1997], focando na previsão de polaridade de adjetivos.

Uma das formas não supervisionadas de classificação leva em consideração aspectos sintáticos e semânticos do texto. [Turney 2002] apresenta uma abordagem de expansão léxica fazendo uso da técnica de *Pointwise Mutual Infomation* (PMI), com o objetivo de calcular a coocorrência de palavras e, com isso, comparar a polaridade de novas palavras com outras já conhecidas. Nesse trabalho, amplamente referenciado por outras pesquisas, o autor compara o conjunto de palavras de Orientação Semântica desconhecida com as palavras *"excellent"* e *"poor"*, representando Orientações Semânticas positiva e negativa, respectivamente. Essas palavras previamente conhecidas utilizadas como base para a expansão do Dicionário são chamadas de palavras semente (*seed words*, em inglês). Como exemplo de trabalhos que utilizam o PMI para a criação e expansão do Dicionário Léxico podemos citar [Becker et al. 2013], [Zhou et al. 2014], [Pinto et al. 2007]. [Pantel and Pennacchiotti 2006], [Duwairi et al. 2015], entre outros.

Outro importante trabalho sobre Mineração de Opiniões, [Taboada et al. 2011] apresenta uma abordagem de Análise de Sentimentos baseada em Léxico combinada com uma verificação manual. Esse trabalho apresenta o SO-CAL (*Semantic Orientation Calculator*), que usa lista de palavras já consolidadas para a geração de dicionários com novas entradas e suas polaridades de forma não supervisionada.

Na linha de estratégias supervisionadas, [Eisenstein 2016] e [Bandhakavi et al. 2016] apresentam outros procedimentos para apoiar a Análise de Sentimentos. O primeiro apresenta uma abordagem usando a técnica de *Naive Bayes* para a classificação dos aspectos e cita problemas de estimativas de palavras e avaliação dos léxicos criados. O segundo faz uma comparação de algumas técnicas de avaliação em 4 conjuntos de dados diferentes, apresentando uma análise quantitativa do mesmo. Abordagens e comparações semelhantes, com algumas modificações no domínio e no idioma do problema abordado, podem ser vistos em [Khoo and Johnkhan 2017], [Asghar et al. 2014] e [Ding et al. 2008].

Sistemas Classificadores, como o criado em [Rodrigues et al. 2016], recebem como entrada um texto e retornam a Orientação Semântica do mesmo. Para esse processo, [Rodrigues et al. 2016] faz uso de um Dicionário próprio, com grande parte das polaridades anotadas manualmente. O processo de classificação, desafios da área e outros exemplos de classificadores são discutidos em [Pang et al. 2002], [Zhou et al. 2014], [Silva et al. 2010], [Guimaraes et al. 2016], entre outros.

Importante destacar, como bem apresentado por [Araújo et al. 2013], que somente o Dicionário Léxico não é capaz de prover uma classificação dos sentimentos eficaz e a simples soma das polaridades das palavras pode apresentar resultados não satisfatórios, levando a uma avaliação incorreta. Grande parte dos classificadores possuem heurísticas que trabalham em conjunto com os dicionários, além de processamentos prévios das mensagens, o que proporciona um melhor resultado das classificações.

De forma geral, essas estratégias heurísticas levam em consideração aspectos gramaticais e sintáticos que tem possuem uma importância na expressão do sentimento, como pontuação, negação, intensificação, capitalização, entre outros. Em contextos específicos, como a classificação de *Tweets*, pode-se levar em consideração a quantidade de *hashtags*, gírias, etc.

A criação de modelos pode ser vista como um problema de otimização. Para essa classe de problemas, podemos fazer uso de estratégias de computação evolucionária, baseadas na teoria da evolução de *Darwin*. Dentre os trabalhos que abordam a Análise de Sentimentos fazendo uso de Estratégias Evolutivas, podem citar [Ferreira et al. 2015], [Vohra and Teraiya 2013], [Haddi et al. 2013] e [Silveira et al. 2014].

Conjuntos de dados previamente avaliados são amplamente utilizados para teste dos Sistemas de Classificação. Esses dados tem sua Orientação Semântica determinadas por especialistas humanos, e servem como entrada para a comparação da saída dos classificadores, ou seja, são dados considerados corretos e consistentes. Principais conjuntos de dados disponíveis para utilização no processo de Análise de Sentimentos podem ser vistos em [Hu and Liu 2004], [Iqbal et al. 2015], [Taboada et al. 2011], entre outros.

4. Abordagem

Nesta seção, serão apresentadas as abordagens de análise do problema de pesquisa e do projeto da solução, com detalhes relevantes de implementação, dados utilizados, bibliotecas de apoio, entre outros.

4.1. Análise do problema

A criação de um classificador de sentimentos para um determinado contexto é um desafio de pesquisa na área de Análise de Sentimentos. Detalhes intrínsecos do domínio do texto, idioma, entre outros, podem ser relevantes para as regras de classificação.

O modelo de classificação, portanto, pode ser descrito como um programa. Podemos abordar essa situação como um problema de otimização, com o objetivo de encontrar um modelo que represente a solução desejada.

A Programação Genética pode auxiliar nesse processo de criação do modelo de classificação. De posse de um conjunto de dados previamente classificados, podemos treinar nossa população de possíveis soluções (indivíduos), avaliando seu *fitness* de acordo com a semelhança com o resultado esperado para determinada entrada.

Por possuir uma forma de atuação paralela, essa abordagem resulta em soluções muito próximas da solução ótima (às vezes encontra a melhor resolução) para problemas complexos.

O indivíduo mais apto (de melhor *fitness*) retornado pelo algoritmo de Programação Genética tem grandes chances de ser um modelo de classificação de sentimentos com bons resultados.

4.2. Projeto da solução

Como explanado na seção 2.2, a Programação Genética pode ser utilizada para a criação de um modelo de solução - um programa - para a resolução de um dado problema. No caso do presente trabalho, queremos encontrar um modelo de classificação de sentimentos em *Tweets*.

O primeiro passo para projetar uma solução de Programação Genética é determinar o conjunto de terminais e funções do modelo. Os terminais serão compostos pelos *Tweets* que serão analisados, bem como por uma constante efêmera, um número real escolhido aleatoriamente entre -3 e 3. A constante efêmera será escolhida de forma aleatória somente no momento de criação do indivíduo, e depois manterá seu valor na árvore.

As funções serão responsáveis por realizar a manipulação dos dados, como o *Tweet* e suas características. A lista das principais funções definidas para a solução pode ser vista na tabela 1.

Para a criação das funções, foi levado em consideração heurísticas apresentadas em trabalhos anteriores sobre o tema e para determinados contextos, como os apresentados em [Araújo et al. 2013], [Rodrigues et al. 2016], [Turney 2002], entre outros.

Função	Retorno
polaritySum(str): float	Soma das polaridades de cada palavra
hashtagPolaritySum(str): float	Soma das polaridades de cada hashtag
emoticonPolaritySum(str): float	Soma das polaridades de cada emoticon
positiveWordsQuantity(str): float	Quantidade de palavras positivas
negativeWordsQuantity(str): float	Quantidade de palavras negativas
hasHashtags(str): bool	Verifica se o *Tweet* possui hashtag
hasEmoticons(str): bool	Verifica se o *Tweet* possui emoticon
removeStopWords(str): str	Remove os *stopwords* do *Tweet*

Table 1. Principais funções utilizadas na Programação Genética

Além das funções citadas na tabela 1, também foram incluídas funções matemáticas como adição, subtração, divisão, multiplicação, logaritmo, raiz quadrada, exponenciação, seno e cosseno.

Outra característica importante a ser definida para o trabalho é a função de *fitness*. Para a solução, o *fitness* é definido como o F1 médio do classificador, calculado utilizando a fórmula apresentada na tabela 3

Terminologia	Descrição
Verdadeiro Positivo (VP)	Modelo retornou Positivo e a classe real é Positivo
Verdadeiro Negativo (VN)	Modelo retornou Negativo e a classe real é Negativo
Verdadeiro Neutro (VNt)	Modelo retornou Neutro e a classe real é Neutro
Falso Positivo (FP)	Modelo retornou Positivo e a classe real não é Positivo
Falso Negativo (FN)	Modelo retornou Negativo e a classe real não é Negativo
Falso Neutro (FNt)	Modelo retornou Neutro e a classe real não é Neutro

Table 2. Dados utilizados para as métricas do modelo

Outras métricas são utilizadas para a verificação da qualidade do classificador: acurácia, precisão e *recall*. Detalhes de cada uma dessas funções são apresentadas na tabela 3. Para todas essas medições, são considerados 6 possibilidades de classificação dos *Tweets*: Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Verdadeiro Neutro (VNt), Falso Positivo (FP), Falso Negativo (FN) e Falso Neutro (FNt), conforme apresentado na tabela 2.

Métrica	Fórmula
Acurácia	(VP + VN + VNt)/ Total de mensagens
Precisão positiva (PP)	VP / Positivos retornados pelo modelo
Recall positivo (RP)	VP / Mensagens positivas
F1 positivo (FP)	2 x (PP * RP) / (PP + RP)
Precisão negativa (PN)	VN / Negativos retornados pelo modelo
Recall negativo (RN)	VN / Mensagens negativas
F1 negativo (FN)	2 x (PN * RN) / (PN + RN)
Precisão neutra (PNt)	VNt / Neutros retornados pelo modelo
Recall neutro (RNt)	VNt / Mensagens neutras
F1 neutro (FNt)	2 x (PNt * RNt) / (PNt + RNt)
Precisão média	(PP + PN + PNt) / 3
Recall médio	(RP + RN + RNt) / 3
F1 médio	(FP + FN + FNt) / 3
F1 médio SemEval	(FP + FN) / 2

Table 3. Métricas utilizadas para avaliar o modelo de classificação

4.3. Bibliotecas

Para apoiar no desenvolvimento da solução do problema, foi utilizada a biblioteca DEAP[1] (*Distributed Evolutionary Algorithms in Python*), escrita na linguagem *Python* e disponível para uso gratuito. Fornece abstrações para a implementação de várias classes de algoritmos evolucionários, como Algoritmos Genéticos, Programação Genética, entre outros. [Fortin et al. 2012]

Especificamente para o contexto de Programação Genética, DEAP fornece funcionalidades para controle de criação das estruturas de árvores, operadores genéticos, parametrização das operações, *logs*, entre outras.

Para a stemização - processo de redução de palavras flexionadas para sua forma raíz - foi utilizada a biblioteca stemming 1.0[2] do *python*.

Para a criação de uma lista de *stopwords* - palavras que podem ser consideradas irrelevantes para a análise do texto - foi utilizada a biblioteca nltk[3].

4.4. Dicionários

Para o presente trabalho, foram utilizados os dicionários de palavras positivas e negativas de [Hu and Liu 2004][4]. Os dicionários fornecem um conjunto de 4783 palavras negativas e 2006 palavras positivas para apoiar no processo de Análise de Sentimentos.

Utilizou-se, também, o dicionário de *emoticons* SentiStrength[5], que fornece 46 *emoticons* positivos e 58 negativos.

A escolha desses dicionários deu-se, principalmente, por terem sido utilizados como base para o SemEval 2014, *task* 9. Ao considerar as mesmas bases disponibilizadas

[1]https://github.com/DEAP/deap

[2]https://pypi.python.org/pypi/stemming/1.0

[3]http://www.nltk.org/

[4]https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html#lexicon

[5]http://sentistrength.wlv.ac.uk/

pela competição, é possível fazer a comparação dos resultados obtidos neste trabalho com os avaliados pelo evento.

4.5. *Datasets*

Há diversas bases de dados anotadas disponíveis na Internet para *download*. O escopo do presente trabalho é a Análise de Sentimentos em *Tweets*, por isso utilizou-se uma base disponibilizada para o evento SemEvaL 2014[6] (*International Workshop on Semantic Evaluation*), uma das principais referências na área de análise semântica.

O evento é dividido por *Tasks*, que possuem objetivos distintos dentro da área de pesquisa. Para este trabalho, utilizou-se a base de dados da *Task 9 - Sentiment Analysis in Twitter*. São disponibilizadas bases de treinamento e de testes para *download*[7] no site do evento.

A base de treinamento aplicada no trabalho possui 9684 *Tweets*, com a seguinte divisão de polaridades: 3640 mensagens positivas, 1458 negativas e 4586 neutras.

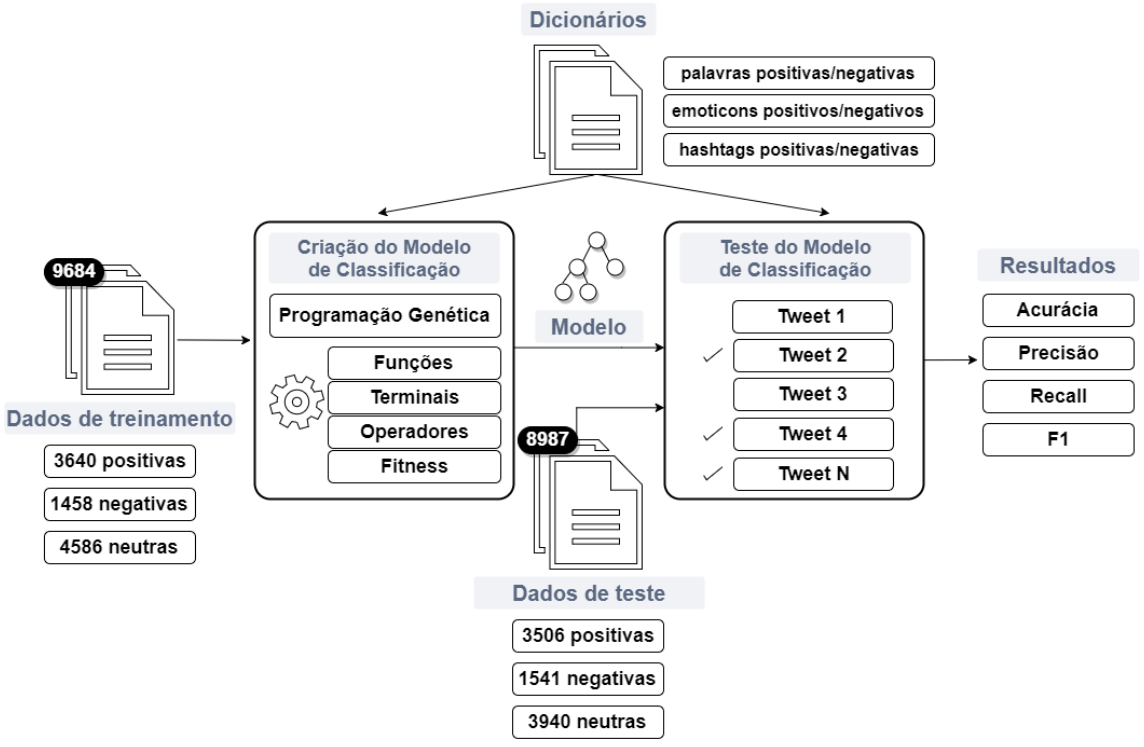

Figure 2. Diagrama simplificado da solução

O evento também disponibiliza uma base de testes, que serve como critério de avaliação e comparação dos trabalhos submetidos para cada *task*. A base fornecida é dividida em 5 categorias, e possui um total de 8987 *Tweets*. Detalhes da divisão e da polaridade das bases podem ser visualizadas na tabela 4.

[6]http://alt.qcri.org/semeval2014/

[7]http://alt.qcri.org/semeval2014/task9/index.php?id=data-and-tools

Base de Dados	*Tweets*
Tweets2013	3813
Tweets2014	1853
Tweets2014Sarcasm	86
SMS2013	2093
LiveJournal2014	1142
Total	8987

Table 4. Bases de testes - Semeval2014

5. Resultados

Nesta seção, serão apresentados os resultados obtidos com os modelos gerados pelo sistema.

5.1. Criação dos modelos

A criação do modelo é feita pelo processo de Programação Genética, tendo como entrada os *Tweets* da base de treinamento utilizada no trabalho e discutida na seção 4. Os processos de Programação Genética encarregam-se da criação e evolução da população de indivíduos, conforme os princípios citados anteriormente.

Alguns parâmetros devem ser definidos para o funcionamento da Programação Genética. Os principais deles são quantidade de gerações, população, taxa de *crossover* e taxa de mutação. Não há uma regra para a definição desses valores, e cada problema deve ter uma abordagem diferente. Para o presente trabalho, os parâmetros definidos para a criação e evolução dos modelos podem ser vistos na tabela 5.

Modelo	População	Gerações	*Crossover*	Mutação	Gerações sem evolução
A	50	500	3.5%	1.5%	150
B	50	600	9.5%	5.5%	300
C	100	500	4.5%	2.5%	500

Table 5. Parâmetros da Programação Genética

Com o objetivo de diminuir a quantidade de tempo de processamento para a criação dos modelos pelo algoritmo de Programação Genética, foi criado um parâmetro que determina a quantidade máxima aceitável de gerações sem evolução no *fitness*, ou seja, a quantidade de ciclos seguidos em que não houve melhoria no melhor indivíduo da população.

Inicialmente, foram criados 3 modelos de classificação de sentimentos, usando as mesmas entradas (base de treinamento) e modificando somente os parâmetros da Programação Genética. Cada um dos modelos foi testado com a base de testes disponibilizada pelo SemEval 2014.

Os modelos criados após o processamento foram (a variável x representa a entrada, nesse caso, os *Tweets*):

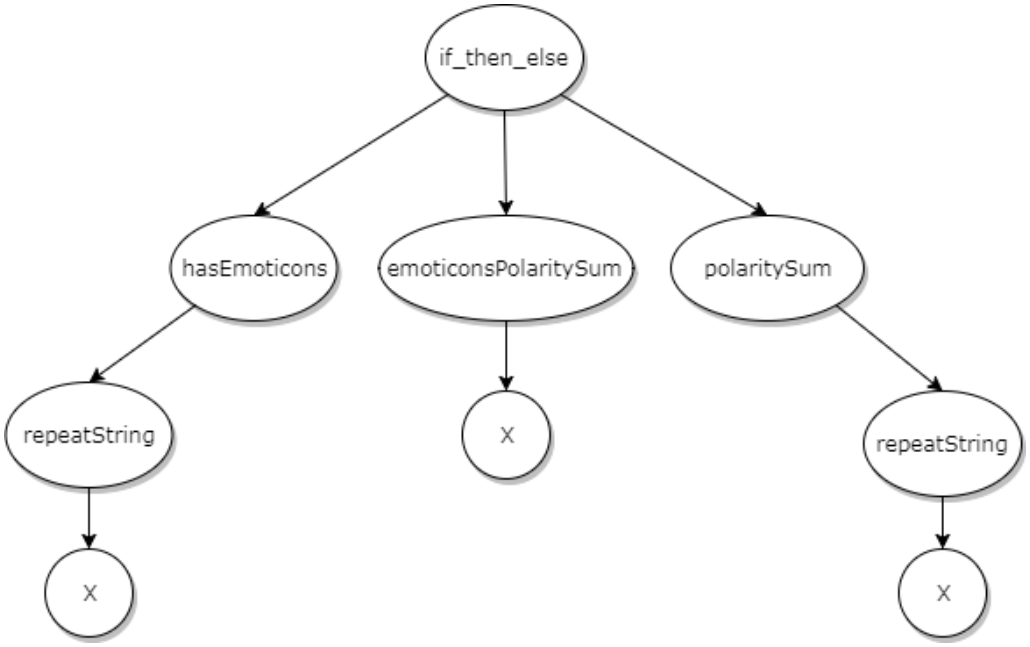

Figure 3. Modelo A

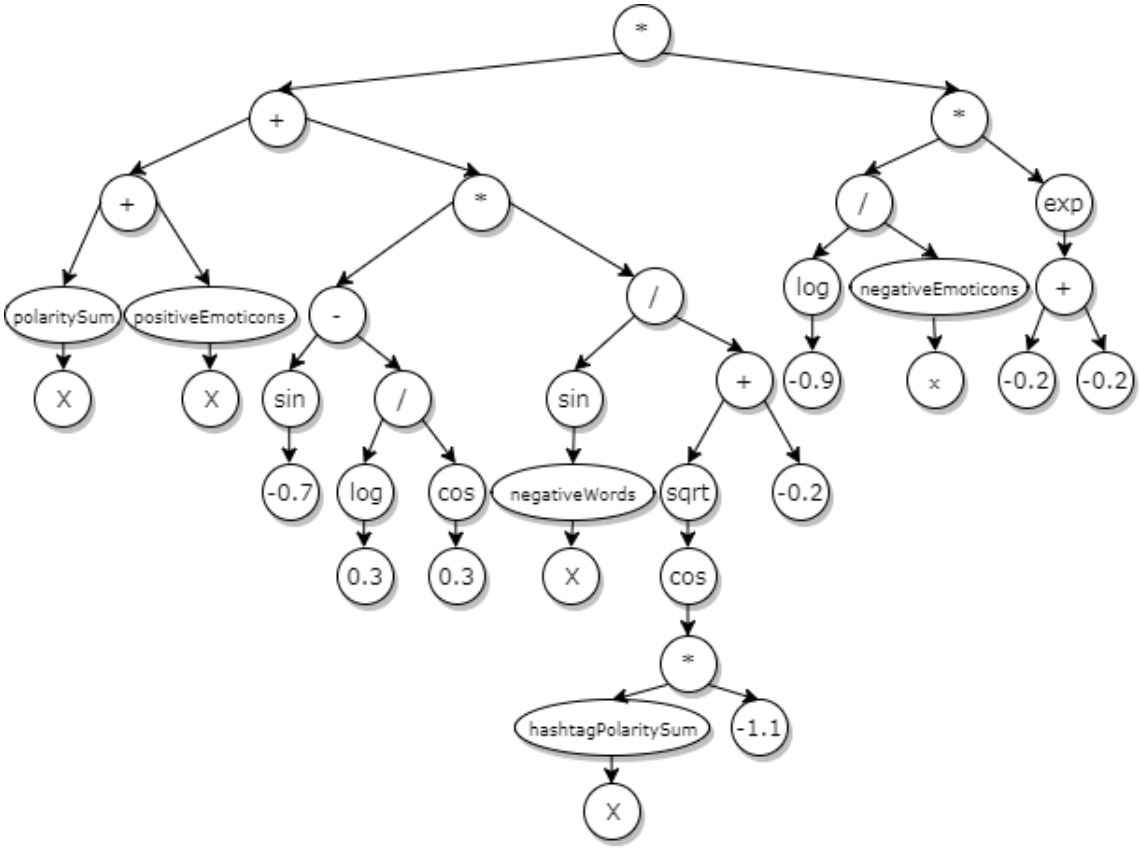

Figure 4. Modelo B

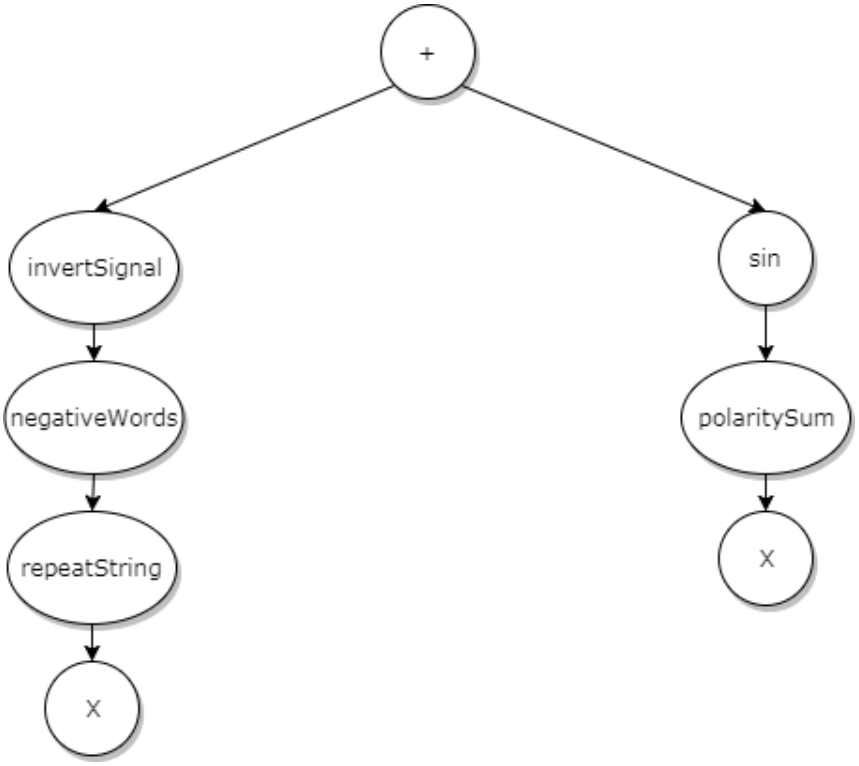

Figure 5. Modelo C

Além dos modelos criados por meio do processo de Programação Genética, foram realizados testes com um modelo de classificador padrão simples, que faz a soma das polaridades das palavras contidas nas mensagens, representado pela função *polaritySum(x)*. O objetivo principal é criar um ponto de comparação entre um modelo simples e os modelos gerados pelo processo proposto neste trabalho, de forma a identificar os possíveis ganhos nos resultados dos novos classificadores.

5.2. Testes dos modelos

Os modelos foram testados com a base de testes disponibilizada pela organização do evento SemEval 2014. Como apresentado na tabela 4, as mensagens são divididas em 5 bases de teste: Twitter2013, Twitter2014, Twitter2014Sarcasm, SMS2013 e LiveJournal2014. Os resultados foram calculados para cada base e, posteriormente, de forma geral, utilizando todas as mensagens.

A tabela 6 apresenta a quantidade de *Tweets* avaliados corretamente por modelo.

Base	***Tweets***	**Avaliados corretamente**											
		Modelo A				**Modelo B**				Modelo C			
		P	N	Nt	T	P	N	Nt	T	P	N	Nt	T
Tweets2013	3813	778	244	1254	2276	773	280	1253	2288	674	294	1261	2229
Tweets2014	1853	425	71	507	1001	422	79	494	995	354	85	503	942
Sarcasm	86	11	2	11	24	11	3	11	25	11	5	11	27
SMS2013	2093	237	120	993	1350	234	135	986	1355	209	136	999	1344
LiveJournal	1142	257	131	312	700	255	154	307	716	233	157	314	704
Todas	8987	1709	561	3080	5350	1695	651	3033	5379	1481	677	3088	5246

Table 6. Avaliações corretas por base

Com relação ao cálculo de outras métricas (apresentadas na tabela 3), especificamente no processamento de F1, foi criada uma segunda média, levando em consideração somente o F1 positivo e negativo, desconsiderando o neutro. Esse foi o critério adotado pela equipe de avaliação do SemEval 2014 (*task* 9, B), e foi aplicado no presente trabalho para que fosse possível comparar os resultados do mesmo com os classificadores submetidos para a competição. O resultado da *task* está disponível para consulta no site do evento [8].

Os testes foram realizados para cada um dos modelos criados, e os resultados podem ser vistos nas tabelas 10, 8 e 9. Precisão, *Recall* e F1 foram calculados para as mensagens Positivas (P), Negativas (N) e Neutras (Nt), além da Média (Avg). Especificamente para F1, foi calculada, também, a média somente das mensagens positivas e negativas (Avg +/-) pelo motivo citado anteriormente. A acurácia (Acc) representa a quantidade de acertos em relação ao total de mensagens.

Base	**Acc**	**Precisão**				***Recall***				**F1**				
		P	N	Nt	Avg	P	N	Nt	Avg	P	N	Nt	Avg	Avg +/-
Tweets2013	0.6	0.7	0.5	0.6	**0.6**	0.5	0.4	0.8	0.6	0.6	0.5	0.6	**0.6**	0.5
Tweets2014	0.5	0.7	0.4	0.5	0.5	0.4	0.4	0.8	0.5	0.5	0.4	0.6	0.5	0.5
Sarcasm	0.3	0.4	0.5	0.2	0.4	0.3	0.1	0.8	0.4	0.4	0.1	0.3	0.3	0.2
SMS2013	0.6	0.5	0.5	0.7	**0.6**	0.5	0.3	0.8	0.5	0.5	0.4	0.8	0.5	0.4
LiveJournal	0.6	0.7	0.7	0.5	**0.6**	0.6	0.4	0.8	**0.6**	0.7	0.5	0.6	**0.6**	**0.6**
Todas	0.6	0.7	0.5	0.6	**0.6**	0.5	0.4	0.8	0.5	0.6	0.4	0.7	0.6	0.5

Table 7. Resultados dos testes do modelo A

[8] https://docs.google.com/spreadsheets/d/1CmDicfEIxRgyoAix9BsVcC3qoRFEq0XDTSnBLVdavu8/

Base	**Acc**	**Precisão**				***Recall***				**F1**				
		P	N	Nt	Avg	P	N	Nt	Avg	P	N	Nt	Avg	Avg +/-
Tweets2013	0.6	0.7	0.5	0.6	**0.6**	0.5	0.5	0.8	**0.6**	0.6	0.5	0.7	**0.6**	0.5
Tweets2014	0.5	0.7	0.3	0.5	0.5	0.4	0.4	0.7	0.5	0.5	0.4	0.6	0.5	0.5
Sarcasm	0.3	0.4	0.4	0.2	0.3	0.3	0.1	0.8	0.4	0.4	0.1	0.3	0.3	0.2
SMS2013	0.6	0.5	0.5	0.7	**0.6**	0.5	0.3	0.8	0.5	0.5	0.4	0.8	**0.6**	0.5
LiveJournal	0.6	0.7	0.6	0.6	**0.6**	0.6	0.5	0.7	**0.6**	0.7	0.6	0.6	**0.6**	**0.6**
Todas	0.6	0.7	0.5	0.6	**0.6**	0.5	0.4	0.8	**0.6**	0.6	0.4	0.7	**0.6**	0.5

Table 8. Resultados dos testes do modelo B

Base	**Acc**	**Precisão**				***Recall***				**F1**				
		P	N	Nt	Avg	P	N	Nt	Avg	P	N	Nt	Avg	Avg +/-
Tweets2013	0.6	0.7	0.5	0.6	**0.6**	0.4	0.5	0.8	**0.6**	0.5	0.5	0.7	**0.6**	0.5
Tweets2014	0.5	0.7	0.3	0.5	0.5	0.4	0.4	0.8	0.5	0.5	0.4	0.6	0.5	0.4
Sarcasm	0.3	0.4	0.6	0.2	0.4	0.3	0.1	0.8	0.4	0.4	0.2	0.3	0.3	0.3
SMS2013	0.6	0.5	0.5	0.7	**0.6**	0.4	0.3	0.8	0.5	0.5	0.4	0.8	0.5	0.4
LiveJournal	0.6	0.7	0.6	0.6	**0.6**	0.5	0.5	0.8	**0.6**	0.6	0.6	0.7	**0.6**	**0.6**
Todas	0.6	0.7	0.5	0.6	**0.6**	0.4	0.4	0.8	0.5	0.5	0.4	0.7	0.5	0.5

Table 9. Resultados dos testes do modelo C

Base	**Acc**	**Precisão**				***Recall***				**F1**				
		P	N	Nt	Avg	P	N	Nt	Avg	P	N	Nt	Avg	Avg +/-
Tweets2013	0.6	0.7	0.5	0.5	**0.6**	0.4	0.4	0.8	0.5	0.5	0.4	0.6	0.5	0.5
Tweets2014	0.5	0.7	0.4	0.4	0.5	0.4	0.3	0.8	0.5	0.5	0.3	0.6	0.5	0.4
Sarcasm	0.3	0.4	0.5	0.2	0.4	0.3	0.1	0.8	0.4	0.3	0.1	0.3	0.2	0.2
SMS2013	0.6	0.5	0.5	0.7	**0.6**	0.4	0.3	0.8	0.5	0.5	0.4	0.7	0.5	0.4
LiveJournal	0.6	0.7	0.7	0.5	**0.6**	0.6	0.4	0.8	**0.6**	0.6	0.5	0.6	**0.6**	**0.6**
Todas	0.6	0.7	0.5	0.6	**0.6**	0.4	0.3	0.8	0.5	0.5	0.4	0.6	0.5	0.5

Table 10. Resultados dos testes do modelo *polaritySum(x)*

Para facilitar a comparação do resultado entre os modelos, a tabela 11 apresenta as médias das métricas calculadas para cada base de teste e, também, para o total de mensagens.

	Base	Acurácia	Precisão (média)	Recall (médio)	F1 (média)	F1 (+/-)
Modelo A	Tweets2013	0.5969	0.5886	0.5552	0.5599	0.5137
	Tweets2014	0.5413	0.5309	0.514	0.4967	0.458
	TwitterSarcasm	0.2791	0.3623	0.4098	0.2597	0.2229
	SMS2013	0.645	0.5898	0.5363	0.5498	0.4472
	LiveJournal	0.613	0.6381	0.5973	0.601	0.5889
	Todas	0.5956	0.5901	0.5456	0.5531	0.4999
Modelo B	Tweets2013	0.6001	0.5815	0.5702	0.5671	0.5193
	Tweets2014	0.537	0.5141	0.5197	0.4929	0.4507
	TwitterSarcasm	0.2907	0.3426	0.4182	0.2772	0.2412
	SMS2013	0.6474	0.583	0.5451	0.5569	0.4549
	LiveJournal	0.627	0.6364	0.6169	0.6189	0.6052
	Todas	0.5985	0.5814	0.5586	0.5605	0.5054
Modelo C	Tweets2013	0.5846	0.5734	0.5623	0.5529	0.5
	Tweets2014	0.5084	0.5001	0.511	0.4704	0.4205
	TwitterSarcasm	0.314	0.3943	0.4348	0.3067	0.2854
	SMS2013	0.6421	0.5705	0.5325	0.5446	0.435
	LiveJournal	0.6165	0.6265	0.6087	0.6079	0.5861
	Todas	0.5837	0.5701	0.5485	0.5451	0.4833
polaritySum(x)	Tweets2013	0.5796	0.5753	0.5354	0.5377	0.4831
	Tweets2014	0.5175	0.5163	0.4955	0.4726	0.4262
	TwitterSarcasm	0.2791	0.3623	0.4098	0.2597	0.2229
	SMS2013	0.6436	0.5859	0.5283	0.5428	0.4358
	LiveJournal	0.6121	0.6419	0.5968	0.5999	0.5845
	Todas	0.583	0.5807	0.5314	0.5367	0.476

Table 11. Resultados médios por modelo

Uma forma de avaliar a eficiência dos modelos gerados pelo processo é comparar com os resultados dos classificadores submetidos para o SemEval 2014, *task* 9. Como estamos fazendo uso da mesma base de treinamento e de testes utilizados no evento, é possível comparar e avaliar a colocação dos modelos gerados no *ranking* dos trabalhos. A tabela 12 apresenta uma comparação dos melhores resultados obtidos pelo presente trabalho em relação aos trabalhos submetidos para o evento.

O evento, como discutido nas seções anteriores, leva em consideração para a avaliação dos classificadores o F1 médio das mensagens positivas e negativas, desconsiderando as mensagens neutras. Apesar disso, os classificadores não são binários, as mensagens neutras somente não são contabilizadas na média final. Como fator de comparação, ainda na tabela 12 são apresentados os valores de F1 médio dos 3 primeiros colocados, separados por base de dados.

Base	Melhor Modelo	Valor (F1 +/-)	Posição SemEval 2014	Top 3 SemEval
Tweets2013	Modelo B	0.5193	43º	1º 0.7212 2º 0.7075 3º 0.7040
Tweets2014	Modelo A	0.458	48º	1º 0.7096 2º 0.7014 3º 0.6995
TwitterSarcasm	Modelo C	0.2854	52º	1º 0.5816 2º 0.5726 3º 0.5650
SMS2013	Modelo B	0.4549	45º	1º 0.7028 2º 0.6768 3º 0.6751
LiveJournal	Modelo B	0.6052	37º	1º 0.7484 2º 0.7446 3º 0.7399

Table 12. Comparação de resultados com o SemEval 2014 *task*9

6. Conclusão

Como podemos perceber, os resultados do melhor classificador para a base *TwitterSarcasm*, que possui sarcasmo em seu conteúdo, tiveram um valor muito baixo - F1 médio (positivo e negativo) de 0.2854. Isso acontece pela dificuldade de identificação dessa figura de linguagem. A inclusão de novas funções e a utilização de bases de treinamento com mais frases contendo sarcasmo com certeza trarão possibilidades de criação de modelos melhores para esse contexto.

Os melhores resultados dos modelos gerados puderam ser observados na base *LiveJournal*, com um F1 médio de 0.6052. Isso se deve ao fato das mensagens dessa base conterem palavras mais comuns e pouco uso de gírias e abreviações. Isso faz com que as palavras sejam encontradas mais facilmente no dicionário, melhorando o resultado final da classificação.

Apesar dos resultados iniciais estarem abaixo dos melhores classificadores apresentados no SemEval 2014, foi possível responder as questões de pesquisa formuladas no capítulo inicial deste trabalho.

Mostramos que é possível a criação automatizada de um modelo de classificação fazendo uso do processo de Programação Genética. Com os parâmetros, funções e modelos atuais, esses classificadores possuem um resultado abaixo do esperado, e podemos afirmar que tem eficiência média na classificação das mensagens de um modo geral. Em relação a outros classificadores, a melhoria de alguns itens pontuais, como novos dicionários e funções, possivelmente trará melhores resultados.

O processo de geração de modelos utilizando a Programação Genética é estocástico. Por isso, ????????

6.1. Melhorias futuras

Buscando melhores resultados nos testes, algumas melhorias podem ser realizadas no processo de criação do modelo de classificação. A primeira delas é a inclusão de novas funções, que poderão ser utilizadas pela Programação Genética em busca de um modelo mais adequado ao contexto de classificação.

Modificações nos parâmetros do algoritmo também podem trazer melhores resultados, como um maior número de indivíduos (população), maior número de gerações e, também, alterações nos parâmetros de probabilidade de *crossover* e mutação.

A busca por dicionários complementares também pode auxiliar na melhoria dos resultados do processo. Especificamente no contexto de *Tweets*, um dicionário consistente de *emoticons* e *hashtags*, por exemplo, podem incrementar os resultados de forma significativa.

Treinar o processo com bases de dados maiores e mais balanceadas também pode auxiliar na busca por melhores resultados. Apesar de implementar a funcionalidade de balanceamento de polaridades, o conjunto de dados de treinamento atual pode ser insuficiente para a criação de um modelo com resultados satisfatórios.

References

Araújo, M., Gonçalves, P., and Benevenuto, F. (2013). Métodos para análise de sentimentos no twitter.

Asghar, M. Z., Khan, A., Ahmad, S., and Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research*, 4(3):181–186.

Bandhakavi, A., Wiratunga, N., Padmanabhan, D., and Massie, S. (2016). Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, pages –.

Becker, L., Erhart, G., Skiba, D., and Matula, V. (2013). Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 333–340.

D'Andrea, A., Ferri, F., Grifoni, P., and Guzzo, T. (2015). Article: Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3):26–33. Published by Foundation of Computer Science (FCS), NY, USA.

Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM.

Duwairi, R. M., Ahmed, N. A., and Al-Rifai, S. Y. (2015). Detecting sentiment embedded in arabic social media - a lexicon-based approach. *Journal of Intelligent and Fuzzy Systems*, 29(1):107–117.

Eisenstein, J. (2016). Unsupervised learning for lexicon-based classification. *CoRR*, abs/1611.06933.

Ferreira, L., Dosciatti, M., Nievola, J. C., and Paraiso, E. C. (2015). Using a genetic algorithm approach to study the impact of imbalanced corpora in sentiment analysis. In *FLAIRS Conference*, pages 163–168.

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., and Gagné, C. (2012). DEAP: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13:2171–2175.

Ghaleb, O. A. M. and Vijendran, A. S. (2016). Survey and analysis of recent sentiment analysis schemes relating to social media. *Indian Journal of Science and Technology*, 9(41).

Graff, M., Tellez, E. S., Escalante, H. J., and Miranda-Jiménez, S. (2017). Semantic genetic programming for sentiment analysis. In *NEO 2015*, pages 43–65. Springer.

Guimaraes, N., Torgo, L., and Figueira, A. (2016). Lexicon expansion system for domain and time oriented sentiment analysis. In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016)*, pages 463–471.

Haddi, E., Liu, X., and Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26 – 32. First International Conference on Information Technology and Quantitative Management.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA. ACM.

Iqbal, M., Karim, A., and Kamiran, F. (2015). Bias-aware lexicon-based sentiment analysis. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, pages 845–850, New York, NY, USA. ACM.

Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083, Prague, Czech Republic. Association for Computational Linguistics.

Khoo, C. S. and Johnkhan, S. B. (2017). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 0(0):0165551517703514.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*, volume 1. MIT press.

Liu, B. (2010). Sentiment analysis: A multifaceted problem. *IEEE Intelligent Systems*, 25(3):76–80.

McPhee, N. F., Poli, R., and Langdon, W. B. (2008). Field guide to genetic programming.

Mohammad, S. M. (2016). Challenges in sentiment analysis. *A Practical Guide to Sentiment Analysis, D. Das, E. Cambria, and S. Bandyopadhyay, Eds. Springer*.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 113–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

PATELLI, A. (2011). Genetic programming techniques for nonlinear systems identification.

Pinto, D., Rosso, P., and Jiménez-Salazar, H. (2007). Upv-si: Word sense induction using self term expansion. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 430–433, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., and Rosa, T. C. (2016). Sentihealth-cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, 85(1):80 – 95.

Sannino, G., De Falco, I., and De Pietro, G. (2015). *Indirect Blood Pressure Evaluation by Means of Genetic Programming*, pages 75–92. Springer International Publishing, Cham.

Silva, M. J., Carvalho, P., Costa, C., and Sarmento, L. (2010). Automatic expansion of a social judgment lexicon for sentiment analysis.

Silveira, B. B., Leitão-Júnior, P. S., Ramada, M. S., and Martins, B. P. (2014). Geração de base de dados para o teste de aplicações de banco de dados pelo emprego da computação evolucionária.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vohra, S. and Teraiya, J. (2013). A comparative study of sentiment analysis techniques. *Journal JIKRCE*, 2(2):313–317.

Zhou, Z., Zhang, X., and Sanderson, M. (2014). *Sentiment Analysis on Twitter through Topic-Based Lexicon Expansion*, pages 98–109. Springer International Publishing, Cham.