

Guilherme Henrique Santos Andreato

**O Uso de Processamento de Linguagem
Natural para a Análise de Sentimentos na Rede
Social Reddit**

Caxias do Sul

2017

Guilherme Henrique Santos Andreatta

O Uso de Processamento de Linguagem Natural para a Análise de Sentimentos na Rede Social Reddit

Projeto de Diplomação submetido ao curso de Bacharelado em Sistemas de Informação do Centro de Computação e Tecnologia da Informação da Universidade de Caxias do Sul, como requisito obrigatório para graduação.

UNIVERSIDADE DE CAXIAS DO SUL
CENTRO DE COMPUTAÇÃO E TECNOLOGIA DA INFORMAÇÃO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Orientador: André Luis Martinotto

Caxias do Sul

2017

Guilherme Henrique Santos Andreata

O Uso de Processamento de Linguagem Natural para a Análise de Sentimentos na Rede Social Reddit

Projeto de Diplomação submetido ao curso
de Bacharelado em Sistemas de Informação
do Centro de Computação e Tecnologia da
Informação da Universidade de Caxias do Sul,
como requisito obrigatório para graduação.

André Luis Martinotto
Orientador

Daniel Luís Notari
Convidado 1

Helena Graziottin Ribeiro
Convidado 2

Caxias do Sul
2017

Resumo

A sociedade tem cada vez mais expressado a sua opinião através de Redes Sociais, sendo que entre essas se destaca o Reddit. De fato, essa é uma das maiores redes sociais no mundo, com mais de 17 milhões de usuários. Nesta, os usuários podem postar *links*, bem como comentários sobre estes, gerando um grande volume de dados que muitas vezes são ignorados.

A identificação de padrões de sentimentos expressos por grupos dessa comunidade, se torna útil visto que a partir dessa avaliação é possível construir ferramentas que podem apoiar decisões do ponto de vista político, econômico, etc. Por exemplo, a partir desta análise, é possível identificar a opinião dos usuários em relação a um candidato em uma eleição ou ainda a aceitação dos consumidores em relação a um novo produto.

Assim, neste trabalho foi desenvolvido um *software* que permite efetuar a análise de sentimentos na rede social Reddit. Esse foi desenvolvido utilizando o método de *Valence Aware Dictionary and sEntiment Reasoner* (VADER), que se encontra implementado no *framework Natural Language Toolkit* (NLTK).

Além disso, utilizou-se o Método de Propagação Dupla, com a escolha de palavras alvos. A partir dessa implementação obteve-se uma assertividade de 59,17% na avaliação de comentários de política e 72,54% na avaliação de comentários sobre filmes.

Palavras-chaves: Reddit. Processamento de Linguagem Natural. Análise de Sentimentos.

Abstract

The society has been increasingly expressing themselves through social networks, which from among those, the one who stand out is Reddit. Indeed, this is one of the biggest social networks in the world with more than 17 millions of users. At this, users can send links, as well as comment on these, generating a big volume of data, which a lot of times get ignored.

The recognition of sentiment patterns expressed by groups of that community make itself useful because from that evaluation, is possible the build tools that can support decisions from an economic point of views, political point of view, etc. For an example, with that analysis, it is possible to identify the users's opinion regarding an election candidate or the costumers's acceptance of a new product.

Therefore, in this work was developed a software capable of performing a sentiment analysis on the Reddit social network. This was developed using the VADER method which is implemented by the NLTK *framework*.

Furthermore, it was utilized the Double Propagation Method with the choosing of a target word. From that implementation we got a 59,17% accuracy evaluating political commentaries and 72,54% accuracy evaluating movie reviews commentaries.

Key-words: Reddit. Natural Language Processing. Sentiment Analysis.

Lista de ilustrações

Figura 1 – Caminhos possíveis de classificação	22
Figura 2 – Caminhos definidos para a classificação pelo Algoritmo de Viterbi . . .	26
Figura 3 – Resultados obtidos para definição do método.	33
Figura 4 – <i>Website Reddit</i> : As flechas demarcadas permitem efetuarmos <i>upvotes</i> ou <i>downvotes</i>	35
Figura 5 – Arquitetura do <i>Crawler</i>	36
Figura 6 – Assertividade da ferramenta desenvolvida com relação a comentários de política.	44
Figura 7 – Assertividade da ferramenta desenvolvida com relação a comentários de política por tamanho.	44
Figura 8 – Assertividade da ferramenta desenvolvida com relação a comentários de filmes.	46
Figura 9 – Assertividade da ferramenta desenvolvida com relação a comentários de filmes por tamanho.	46

Lista de tabelas

Tabela 1	–	Tabela de Probabilidades de Associação	23
Tabela 2	–	Tabela de Probabilidade de Transição	25
Tabela 3	–	<i>Training Set</i>	27
Tabela 4	–	Tabela de Palavras e Probabilidades.	29
Tabela 5	–	Tabela de Probabilidades - <i>Laplace smoothing</i>	30

Lista de acrônimos

NLP	<i>Natural Language Processing</i>
NLTK	<i>Natural Language Toolkit</i>
MaxEnt	<i>Maximum Entropy</i>
VADER	<i>Valence Aware Dictionary and sEntiment Reasoner</i>
JSON	<i>JavaScript Object Notation</i>
POJO	<i>Plain Old Java Objects</i>
SVM	<i>Support Vector Machines</i>
API	<i>Application Programming Interface</i>

Sumário

1	INTRODUÇÃO	17
1.1	Objetivos do Trabalho	18
1.2	Estrutura do Trabalho	18
2	PROCESSAMENTO DE LINGUAGEM NATURAL	19
2.1	Linguística	19
2.2	Métodos de Processamento de Linguagem Natural	20
2.2.1	Método Simbólico	20
2.2.2	Método Estatístico	22
3	MÉTODOS ESTATÍSTICOS E SIMBÓLICOS APLICADOS NA ANÁLISE DE SENTIMENTOS	27
3.1	Método de Naive Bayes	27
3.2	Método de <i>VADER</i>	30
3.2.1	Definição de Método para Identificação de Sentimentos	32
3.2.2	NLTK - <i>Natural Language Toolkit</i>	33
4	CRIAÇÃO DE BASE DE DADOS	35
4.1	Rede Social Reddit	35
4.2	Extração de Dados	36
4.3	Validação da Implementação	37
4.4	Método de Propagação Dupla	37
4.4.1	Adição de Palavras e Alvos ao Dicionário	38
4.4.2	Atribuição de Pontuação às Novas Palavras	40
4.4.3	Textos utilizados para o método de Propagação Dupla	40
4.5	Restrição de Palavras Alvos	42
5	RESULTADOS OBTIDOS	43
5.1	Comentários Políticos	43
5.2	Comentários sobre Filmes	45
6	CONCLUSÃO FINAL	47
6.1	Trabalhos Futuros	47
	REFERÊNCIAS	49

1 Introdução

A linguagem é a forma como nós nos comunicamos, seja ela escrita ou falada. De fato, a linguagem é a forma como expressamos nossas idéias, sentimentos e experiências. O Processamento de Linguagem Natural é o termo que é utilizado para descrever um software ou componente de hardware que tem como função analisar a linguagem escrita ou falada (JACKSON; MOULINIER, 2007).

Existem duas abordagens para o Processamento de Linguagem Natural, sendo que a primeira delas é chamada de simbólica (ou racionalista) e a outra de empírica (ou estatística). A primeira abordagem consiste em uma série de regras para a manipulação de símbolos, como as regras gramaticais, que permitem identificar se uma frase está malformada ou não. A abordagem empírica está centrada na análise estatística da linguagem através de uma grande quantidade de texto, como por exemplo, a utilização de modelos de Markov para reconhecer padrões na escrita (JACKSON; MOULINIER, 2007).

Existem diversos *frameworks open source* que facilitam o desenvolvimento de *softwares* para o Processamento de Linguagem Natural, sendo que dentre esses destacam-se o *Stanford's Core NLP Suite* (Stanford CoreNLP, 2017), *Natural Language Toolkit* (Natural Language Toolkit, 2017), *Apache OpenNLP* (Apache OpenNLP, 2017) e *Spacy* (Spacy, 2017). Esses *frameworks* nos permitem, entre outras coisas, efetuar análise de sentimentos, identificar tópicos e conteúdos.

A rede social Reddit é o vigésimo terceiro *website* mais acessado na Internet e o sétimo mais acessado nos Estados Unidos da América (Alexa, 2017). Através deste *website*, seus usuários podem criar ou se inscrever em comunidades, também conhecidas como *subreddits*. Uma vez que as comunidades são criadas pelos próprios usuários, podemos encontrar comunidades sobre todos os assuntos, sejam notícias do mundo, comunidades partidárias ou de pessoas de uma mesma localidade, etc.

Nestas comunidades é possível visualizar e comentar *links* enviados por outros usuários. Além disso, o usuário pode efetuar um voto de forma positiva, caso acredite que aquele *link* é útil para a comunidade, ou um voto negativo em caso contrário. Uma vez que os próprios usuários podem submeter *links*, os eventos e as notícias de todo o mundo são reportados no *website*, como exemplo, pode-se citar as eleições ocorridas no ano de 2016 nos Estados Unidos e o tiroteio ocorrido em Paris em 15 de Novembro de 2015.

Dentro deste contexto, neste trabalho foi desenvolvido um software que permitiu realizar a análise de comentários do *website* Reddit. Mais especificamente, os comentários do Reddit foram analisados com o objetivo de identificar padrões de sentimentos dos usuários, ou seja, determinar se a opinião expressada com relação a um determinado tópico

é neutra, positiva ou negativa.

1.1 Objetivos do Trabalho

Este trabalho apresentou como objetivo a análise dos comentários disponíveis no *website* Reddit, identificando padrões de sentimentos entre os usuários de suas comunidades. De forma a atingir o objetivo principal desse trabalho, os seguintes objetivos específicos foram realizados:

- Construção de uma base de dados a partir do *website* Reddit.
- Desenvolvimento de uma ferramenta para o Processamento Natural de Linguagem através de *frameworks* já existentes.
- Processamento da base de dados criada utilizando-se a ferramenta desenvolvida.

1.2 Estrutura do Trabalho

O presente trabalho está estruturado da seguinte forma:

- No Capítulo 2 é apresentada uma introdução ao Processamento de Linguagem Natural. Além disso, são apresentados exemplos de métodos estatísticos e simbólicos para o processamento de linguagem natural em geral.
- No Capítulo 3 é apresentada a análise de sentimentos. Além disso, é feita uma breve descrição de métodos estatísticos e simbólicos que podem ser aplicados para a análise de sentimentos. Após, é apresentado o método VADER, que é o método que foi utilizado neste trabalho. Por fim, é apresentada a biblioteca NLTK que implementa o método de VADER e que foi a biblioteca utilizada para o desenvolvimento do trabalho.
- O Capítulo 4 tem o objetivo de descrever a base de dados criada para o desenvolvimento do trabalho. Neste Capítulo é descrita a rede Reddit e os tópicos selecionados para a criação da base. Além disso, é descrita a implementação desenvolvida para a extração dos comentários destes tópicos.
- No Capítulo 5 são descritos os resultados obtidos através de testes realizados na implementação e também a análise destes resultados.
- No Capítulo 6 temos as conclusões do trabalho, bem sugestões de trabalhos futuros.

2 Processamento de Linguagem Natural

O objetivo da área de Processamento de Linguagem Natural é analisar a linguagem natural, ou seja, a linguagem utilizada pelo seres humanos seja ela escrita ou falada (MANNING; SCHÜTZE, 1999). O Processamento de Linguagem Natural é uma área antiga, sendo anterior a invenção dos computadores modernos. De fato, sua primeira grande aplicação foi um dicionário desenvolvido no Birkbeck College em Londres no ano de 1948. Por ser uma área complexa, seus primeiros trabalhos foram notavelmente falhos, o que causou uma certa hostilidade por parte das agências fomentadoras de pesquisas.

Os primeiros pesquisadores eram muitas vezes bilíngues, como por exemplo, nativos alemães que imigraram para os Estados Unidos. Acreditava-se que pelo fato desses terem conhecimento de ambas as línguas, Inglês e Alemão, eles teriam capacidade de desenvolver programas de computadores que efetuariam a tradução de modo satisfatório. Uma vez que esses encontraram muitas dificuldades, ficou claro que o maior problema não era o conhecimento das línguas, e sim como expressar esse conhecimento na forma de um programa de computador (HANCOX, 2017).

Para que um computador seja capaz de interpretar uma língua, primariamente, precisamos compreender como nós efetuamos essa interpretação. Por isso, uma parte considerável do Processamento de Linguagem Natural está apoiado na área de Linguística.

2.1 Linguística

O objetivo da Linguística é compreender a forma como os seres humanos adquirem, produzem e entendem as diversas línguas, ou seja, a forma como conversamos, a nossa escrita e outras mídias de comunicação (MANNING; SCHÜTZE, 1999).

Na linguagem tanto escrita, como na falada, existem regras gramaticais que são utilizadas para estruturar as expressões. Uma série de dificuldades no Processamento de Linguagem Natural são ocasionadas pelo fato de que as pessoas constantemente mudam essas regras para satisfazerem suas necessidades de comunicação (MANNING; SCHÜTZE, 1999). Uma vez que as regras são constantemente modificadas pelo locutor, torna-se extremamente difícil a criação de um software ou hardware que efetue a interpretação de uma língua.

2.2 Métodos de Processamento de Linguagem Natural

O *Natural Language Processing* (NLP) tem como objetivo a execução de diferentes tarefas, como por exemplo, a categorização de documentos, a tradução e a geração de textos a partir de um banco de dados, etc. Podemos citar duas classes de métodos para a execução deste tipo de tarefas, que são os métodos simbólicos e os métodos estatísticos.

Nos final dos anos 50 e 60, existiam excelentes métodos estatísticos, que foram desenvolvidos durante a segunda guerra mundial, para a solução de problemas Linguísticos (SHANNON; WEAVER, 1948). Porém, no ano de 1957, Chomsky publicou o trabalho intitulado de “*Syntactic Structures*” onde descreve a teoria da gramática gerativa, que é uma teoria que considera a gramática como um conjunto de regras. Essa abordagem através de um conjunto de regras, ao invés de um modelo estatístico, entra em conflito com os trabalhos anteriores, criando duas comunidades no campo da Linguística. Como reflexo dessas duas comunidades, a área de NLP que crescia em paralelo, também foi dividida em duas áreas. A primeira dessas áreas que fazia uso de métodos baseados em regras (simbólica) e a segunda que fazia o uso de métodos quantitativos (estatística).

Nesta seção será apresentado um exemplo de um método simbólico e de um método estatístico. Destaca-se que a descrição realizada nesta seção apresenta como objetivo, apenas diferenciar ambas as classes de métodos, através de seus requisitos e forma de execução. Destaca-se ainda que os métodos apresentados nesta seção não são utilizados na análise de sentimentos, sendo que os métodos específicos para essa análise serão descritos no Capítulo 3.

2.2.1 Método Simbólico

O método simbólico ou racionalista está baseado no campo da Linguística e faz o uso da manipulação dos símbolos, significados e das regras de um texto. Um exemplo simples de um método simbólico é o método de Brill (BRILL, 1992), que é utilizado para a análise léxica, ou seja, identificar a classe de uma palavra em um texto. Por exemplo, no método de Brill a frase “João pintou a casa de branco”, será separada em palavras que serão classificadas através de um dicionário pré-definido, como:

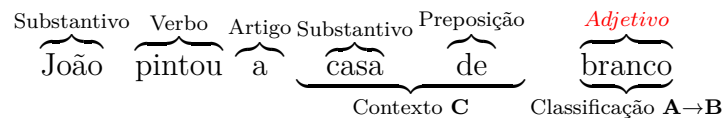
Palavra	João	pintou	a	casa	de	branco
Classificação:		Verbo	Artigo	Substantivo	Preposição	Adjetivo

Observa-se que algumas palavras não foram identificadas, como “João”, ou classificadas de forma incorreta, como “branco”. Desta forma, o método de Brill utiliza-se ainda de duas outras regras para a classificação. A primeira dessas regras classifica todas as palavras desconhecidas que iniciam com uma letra em maiúscula como substantivos, por exemplo, a palavra “João”. Já a segunda regra, atribui para a palavra desconhecida a mesma classificação de outras palavras que terminam com as mesmas três letras. Por exemplo, supondo que a palavra “pintou” não fosse encontrada no dicionário, essa seria associada a outras palavras terminadas com o sufixo “tou”, ou seja, essa seria classificada como verbo.

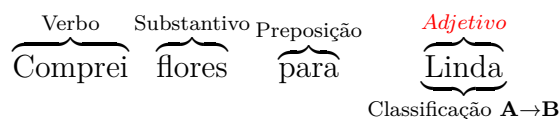
Palavra	João	pintou	a	casa	de	branco
Classificação:	Substantivo	Verbo	Artigo	Substantivo	Preposição	Adjetivo

Após essa classificação inicial, o método executa o seguinte conjunto de regras, ou ainda, regras derivadas dessas:

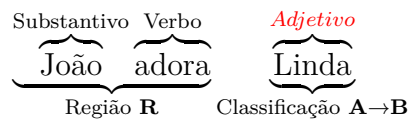
- Se uma palavra tem a classificação **A** e está no contexto **C** então a sua classificação deverá ser mudada para **B**. Por exemplo, se uma palavra **A** (branco no exemplo) é um adjetivo e uma das duas palavras anteriores é uma preposição (“de” no contexto **C**), mude a sua classificação para um substantivo (classificação **B**).



- Se uma palavra tem a classificação **A** e tem uma propriedade **P** então a sua classificação deverá ser alterada para **B**. Por exemplo, se uma palavra **A** (“Linda”) foi classificada como um adjetivo e é iniciada com uma letra maiúscula (propriedade **P**), sua classificação deverá ser alterada para substantivo (classificação **B**).



- Se uma palavra tem a classificação **A** e uma palavra com a propriedade **P** está na região **R**, sua classificação deverá ser **B**. Por exemplo, se uma das duas palavras anteriores à palavra “Linda” (“João adora” na região **R**) iniciam com letra maiúscula (propriedade **P**), sua classificação deverá ser alterada para substantivo (classificação **B**).



2.2.2 Método Estatístico

Um método estatístico utiliza-se de uma grande quantidade de texto, procurando por padrões e associações a modelos, sendo que esses padrões podem ou não estar relacionados com regras sintáticas ou semânticas (MANNING; SCHÜTZE, 1999).

Os métodos estatísticos baseia-se na utilização de um sistema de aprendizado supervisionado, ou seja, a classificação é feita a partir de um conjunto de dados já classificado, que é chamado de *training set*. Um exemplo de método estatístico é a utilização de Modelos de Markov com a aplicação do algoritmo de Viterbi (MANNING; SCHÜTZE, 1999).

Em um Modelo de Markov, a classificação da frase “João comprou um carro” é feita a partir de um *training set* que pode, por exemplo, ser composto por textos retirados de *web-sites*, sendo que as palavras destes textos já devem estar classificadas. A partir deste *training set*, as palavras “João”, “comprou” e “carro” seriam classificadas como substantivo, verbo e substantivo, respectivamente. Já a palavra “um” apresenta uma ambiguidade, uma vez que essa pode ser classificada como um artigo (ART), ou um substantivo (SM) ou um pronome (PRO). A Figura 1 ilustra o conjunto de possibilidades criadas pelo classificador a geração de uma classificação completa da frase.

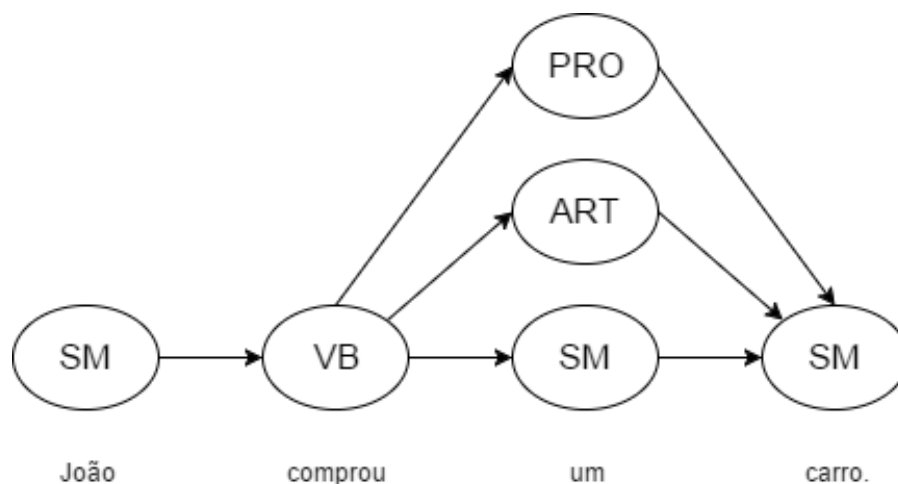


Figura 1 – Caminhos possíveis de classificação

A idéia central da utilização de Modelos de Markov é escolher, entre os caminhos possíveis (Figura 1), o caminho de maior probabilidade. Para tanto, se faz necessário calcular a probabilidade de todos os caminhos através de um Modelo de Markov. Após,

utiliza-se o Algoritmo de Viterbi para definir qual o caminho com maior probabilidade (MANNING; SCHÜTZE, 1999).

O Modelo de Markov irá utilizar-se do *training set* para inferir a classificação da palavra “um”. Por exemplo, considerando-se um *training set* hipotético com as seguintes características:

- 10000 substantivos, onde 150 são a palavra “um”, 10 são a palavra “João” e 50 são a palavra “carro”.
- 20000 artigos, onde 500 são a palavra “um”.
- 12000 verbos, onde 50 são a palavra “comprou”.
- 15000 pronomes, onde 50 são a palavra “um”.

Neste caso, a probabilidade da palavra “um” ser um substantivo é dada pela Equação 2.1, uma vez que no *training set* temos 150 instâncias da palavra “um” classificadas como substantivo e um total de 10000 substantivos. Ou seja, a probabilidade de “um” ser um substantivo é de 0,015. A Equação 2.1 também é aplicada para as demais possíveis classes da palavra “um”, neste caso, pronome ou artigo. Por exemplo, a probabilidade da palavra “um” ser um pronome seria 0,0033 e a probabilidade da palavra “um” ser um artigo seria 0,025.

$$P(\text{palavra}|\text{classe}) = \frac{C(\text{classe}, \text{palavra})}{C(\text{classe})} \quad (2.1)$$

$$P(\text{um}|SM) = \frac{C(SM, \text{um})}{C(SM)} = \frac{150}{10000} = 0,015.$$

O cálculo de probabilidade é realizado para todas as palavras da frase que está sendo classificada. Na Tabela 1 tem-se os resultados obtidos para todas as palavras da frase “João comprou um carro”.

	João	comprou	um	carro
Substantivo	0.001	0	0.015	0.005
Verbo	0	0.0042	0	0
Artigo	0	0	0.025	0
Pronome	0	0	0.0033	0

Tabela 1 – Tabela de Probabilidades de Associação

Além da probabilidade de associação a uma determinada classe, é calculada a probabilidade de transição de uma classe para a outra. Neste caso, vamos considerar que o nosso *training set* hipotético apresenta as seguintes características:

- De 20000 frases, 2500 iniciam com um substantivo, 5000 iniciam com um verbo, 5000 iniciam com um artigo e 5000 iniciam com um pronome.
- De 10000 substantivos, os 10000 são seguidos por verbos.
- De 12000 verbos, 3000 são seguidos por um substantivo, 2000 são seguidos por um outro verbo, 5000 são seguidos por um artigo e 2000 são seguidos por um pronome.
- De 20000 artigos, os 20000 são seguidos por um substantivo.
- De 15000 pronomes, 10000 são seguidos por um substantivo e 5000 são seguidos por um verbo.

A probabilidade de transição de um verbo para um substantivo é dada pela Equação 2.2, uma vez que no *training set* tem-se 12000 verbos, os quais 3000 são seguidos por um substantivo.

$$P(transicao|classe) = \frac{C(classe, transicao)}{C(classe)} \quad (2.2)$$

$$P(SM|VB) = \frac{C(VB, SM)}{C(VB)} = \frac{3000}{12000} = 0,25$$

Da mesma forma, a probabilidade de transição é calculada para todas as demais classes. Por exemplo, a probabilidade de transição de um verbo para outro verbo é de 0,17, de um verbo para um artigo é de 0,42 e de um verbo para um pronome é de 0,17. A Equação 2.2 é utilizada também para o cálculo da probabilidade da frase iniciar com determinada classe. A Tabela 2 tem-se a probabilidade de transição para todas as classes do *training set* de exemplo.

	Substantivo	Verbo	Artigo	Pronome
Início	0.125	0.25	0.25	0.25
Substantivo	0.0	1.0	0.0	0.0
Verbo	0.25	0.17	0.42	0.17
Artigo	1.0	0.0	0.0	0.0
Pronome	0.67	0.33	0.0	0.0

Tabela 2 – Tabela de Probabilidade de Transição

A partir das probabilidades calculadas através do Modelo de Markov, é utilizado o algoritmo de Viterbi para determinar o caminho mais provável. O caminho mais provável é obtido através da Equação 2.3, sendo que essa é aplicada a todas as palavras da frase. Na Equação 2.3, os termos v_t , v_{t-1} , a_{ij} e $b_j(o_t)$ correspondem, respectivamente, o caminho mais provável atual, o caminho mais provável anterior, a probabilidade de transição e a probabilidade de associação. Por exemplo, para a palavra “João”, tem-se que v_{t-1} é igual a 1 (visto que essa é a primeira palavra da frase); a_{ij} (probabilidade de transição entre “Início” e um substantivo) é igual a 0,125 (Tabela 2); e $b_j(o_t)$ (probabilidade de associação da palavra João com um substantivo) é igual a 0,001 (Tabela 1). Desta forma, tem-se que o valor de v_t para a palavra João é:

$$v_t(j) = v_{t-1}a_{ij}b_j(o_t) \quad (2.3)$$

$$v_t(j) = 1 * 0,125 * 0,001 = 0,000125. \quad (2.4)$$

Já para a palavra “comprou” tem-se 0,000125 que é o caminho mais provável anterior (Equação 2.3). Já os valores 1 e 0,0042 são as probabilidades de transição (Tabela 2) e associação (Tabela 1) respectivamente.

$$v_t(j) = 0,000125 * 1 * 0,0042 = 0,000000525. \quad (2.5)$$

Após efetuar o cálculo de probabilidade de todos os caminhos, é escolhido o caminho que tem a maior probabilidade, sendo que neste caso o caminho que apresenta a maior probabilidade é o que possui a palavra “um” como artigo. De fato, esse possui uma probabilidade de 0,0000000055125 ($0,000000525 * 0,42 * 0,025$), já o que apresenta a palavra “um” como pronome possui uma probabilidade de 0,00000000294525 ($0,000000525 * 0,17 * 0,0033$), enquanto o que apresenta a palavra “um” como substantivo possui uma probabilidade de 0,00000000196875 ($0,000000525 * 0,25 * 0,015$).

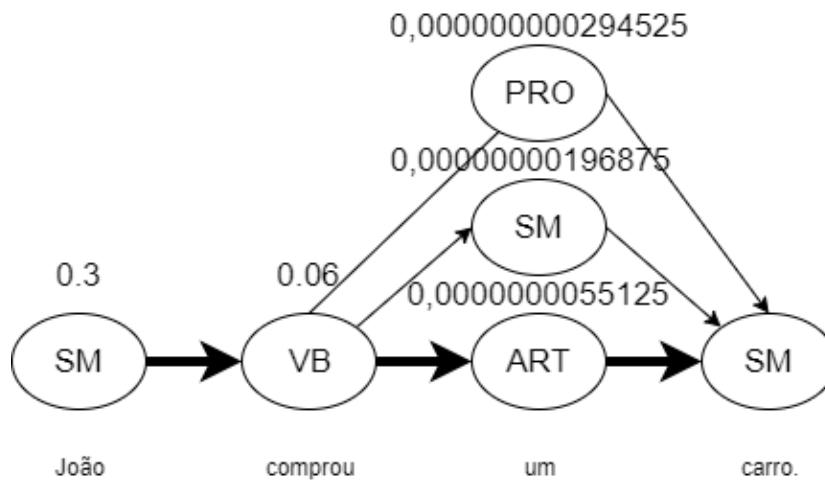


Figura 2 – Caminhos definidos para a classificação pelo Algoritmo de Viterbi

3 Métodos estatísticos e simbólicos aplicados na análise de sentimentos

Antigamente, para sabermos a opinião de outras pessoas sobre um determinado produto, tínhamos que perguntar diretamente a essas pessoas. Com a popularização da Internet e também de redes sociais, milhares de pessoas compartilham as suas opiniões sobre produtos, política, serviços e demais assuntos. Porém, muitas vezes essas opiniões acabam por ser esquecidas devido a dificuldade em analisar uma grande quantidade de textos. De fato, uma das maiores dificuldades reside em como obter a opinião geral das pessoas sobre determinado produto em uma seção de comentários com mais de 1000 opiniões diferentes. Dentro dessa contexto, a análise de sentimentos, tem como função identificar e quantificar os sentimentos expressos através de textos.

Neste capítulo serão descritos um método estatístico (*Naive Bayes*) e um método simbólico (VADER), que podem ser aplicados na análise de sentimentos. Outros possíveis métodos estatísticos para a análise de sentimentos são *Support Vector Machines* (SVM) (HEARST, 1998) e *Maximum Entropy* (MaxEnt) (BERGER; PIETRA; PIETRA, 1996), os quais possuem performance similar ao Naive Bayes (PANG; LEE; VAITHYANATHAN, 2002). Ambos métodos serão descritos para o uso da língua Inglesa, visto que o *Website* analisado (Reddit) possui a maioria de seus comentários em língua Inglesa. Além disso, não foram encontrados métodos que façam uso da língua Portuguesa com similar precisão.

3.1 Método de Naive Bayes

O Naive Bayes é um método estatístico para a classificação e que pode ser utilizado para a análise de sentimento. Esse faz o uso do teorema de Bayes e um *training set* para inferir a classificação de uma frase. Por exemplo, considerando que se quer determinar se a frase “*This place is great*” demonstra um sentimento negativo ou positivo.

Texto	Categoria
The food was great	Positiva
They are horrible!	Negativa
I love the food here	Positiva
This place is wonderful	Positiva
Forgettable experience	Negativa

Tabela 3 – *Training Set*

A partir do *training set* hipotético (Tabela 3) será calculada a probabilidade da frase “*This place is great*” ser positiva e também de ser negativa sendo que a partir dessas

duas possibilidades, será escolhida a de maior probabilidade.

Para calcular a probabilidade da frase “*This place is great*” pertencer a cada categoria é utilizado o teorema de Bayes (MANNING; SCHÜTZE, 1999), através da Equação 3.1. Na Equação 3.1, o termo $P(c|d)$ corresponde a probabilidade da frase **d** pertencer a classe **c**, ou seja, a probabilidade de “*This place is great*” ser uma frase positiva ou negativa. O Termo $P(d|c)$ é a probabilidade da classe **c** ser a frase **d**, ou seja, dentre todas as frases negativas ou positivas, a probabilidade de uma das frases ser “*This place is great*”. Já $P(c)$ é a probabilidade de frases negativas ou positivas aparecerem em nosso *training set*. E, por fim, $P(d)$ é a probabilidade da frase “*This place is great*” aparecer em nosso *training set*.

$$P(c|d) = \frac{P(d|c) \times P(c)}{P(d)} \quad (3.1)$$

$$P(Negativa|This\ place\ is\ great) = \frac{P(This\ place\ is\ great|Negativa) \times P(Negativa)}{This\ place\ is\ great} \quad (3.2)$$

$$P(Positiva|This\ place\ is\ great) = \frac{P(This\ place\ is\ great|Positiva) \times P(Positiva)}{This\ place\ is\ great} \quad (3.3)$$

Uma vez que as Equações 3.2 e 3.3 apresentam $P(d)$ como divisor (“*This place is great*”), essas podem ser simplificadas resultando nas Equações 3.4 e 3.5.

$$P(Negativa|This\ place\ is\ great) = P(This\ place\ is\ great|Negativa) \times P(Negativa) \quad (3.4)$$

$$P(Positiva|This\ place\ is\ great) = P(This\ place\ is\ great|Positiva) \times P(Positiva) \quad (3.5)$$

Nas Equações 3.4 e 3.5, os termos $P(Positiva)$ e $P(Negativa)$ são definidos pela frequência que frases positivas e negativas aparecem no *training set*, sendo determinados através das Equações 3.6 e 3.7. Como pode ser observado, os valores de $P(positiva)$ e $P(negativa)$ correspondem a 0,6 e 0,4 respectivamente, uma vez que em nosso *training set* tem-se um total de 5 frases onde 3 dessas são positivas (Equação 3.6) e as outras 2 são negativas (Equação 3.7).

$$P(Positiva) = \frac{3}{5} = 0,6 \quad (3.6)$$

$$P(Negativa) = \frac{2}{5} = 0,4 \quad (3.7)$$

Uma vez que a frase “*This place is great*” não existe por completo no *training set*, tem-se que o termo $P(d|c)$ da Equação 3.4 e 3.5 é igual a zero (0), impossibilitando o cálculo de probabilidade para essa frase. Neste caso, se faz o uso do *Naive Bayes*, o qual passa a considerar todas as palavras ao invés da frase completa. O uso do Naive Bayes elimina o problema de frases que não se encontram no *training set*. Neste caso, é considerado somente a frequência que cada palavra aparece em uma frase positiva e em uma negativa. Portanto, o termo $P(\textit{This place is great}|\textit{Positiva})$ da Equação 3.5 é dado pela Equação 3.8.

$$\begin{aligned} P(\textit{This place is great}|\textit{Positiva}) &= P(\textit{This}|\textit{Positiva}) \times P(\textit{place}|\textit{Positiva}) \\ &\times P(\textit{is}|\textit{Positiva}) \times P(\textit{great}|\textit{Positiva}) \end{aligned} \quad (3.8)$$

A partir da Equação 3.8 é necessário calcular os termos $P(\textit{This}|\textit{Positiva})$, $P(\textit{place}|\textit{Positiva})$, $P(\textit{is}|\textit{Positiva})$, $P(\textit{great}|\textit{Positiva})$. O termo $P(\textit{This}|\textit{Positiva})$ é calculado pela razão entre a quantidade de vezes que a palavra *This* encontra-se em uma frase classificada como positiva no *training set*, e o total de palavras que se encontram em frases classificadas como positiva (Equação 3.9).

$$P(\textit{This}|\textit{Positiva}) = \frac{1}{13} \quad (3.9)$$

Da mesma forma, a Equação 3.9 deve ser aplicada para as demais palavras da frase “*This place is great*”, obtendo-se os valores apresentados na Tabela 4.

Palavra	Positiva	Negativa
This	$\frac{1}{13}$	$\frac{0}{5}$
place	$\frac{1}{13}$	$\frac{0}{5}$
is	$\frac{1}{13}$	$\frac{0}{5}$
great	$\frac{1}{13}$	$\frac{0}{5}$

Tabela 4 – Tabela de Palavras e Probabilidades.

Uma vez que algumas palavras não encontram-se no *training set* essas acabam zerando o resultado final da multiplicação das probabilidades (Equação 3.8). De modo, a evitar que uma única palavra invalide a frase por completo é utilizado o método *Laplace smoothing* (MANNING; RAGHAVAN; SCHÜTZ, 2008). Neste, é somado 1 a cada palavra da frase. Já ao total de palavras positivas é somada a quantidade de palavras distintas do *training set* (16). Aplicando o *Laplace smoothing* para os valores apresentados na Tabela 4 são obtidos os valores que se encontram na Tabela 5.

Palavra	Positiva	Negativa
This	$\frac{1+1}{13+16}$	$\frac{0+1}{5+16}$
place	$\frac{1+1}{13+16}$	$\frac{0+1}{5+16}$
is	$\frac{1+1}{13+16}$	$\frac{0+1}{5+16}$
great	$\frac{1+1}{13+16}$	$\frac{0+1}{5+16}$

Tabela 5 – Tabela de Probabilidades - *Laplace smoothing*.

Utilizando as probabilidades que são apresentadas na Tabela 5 na Equação 3.8 tem-se:

$$P(\text{Positiva} | \text{This place is great}) = \frac{1+1}{13+16} \times \frac{1+1}{13+16} \times \frac{1+1}{13+16} \times \frac{1+1}{13+16} = 0,000023. \quad (3.10)$$

Uma vez que o termo $P(\text{Positiva} | \text{This place is great})$ encontra-se definido, esse pode ser utilizado na Equação 3.4, onde o termo $P(\text{Positiva})$ é igual a 0,6 (Equação 3.6). Neste caso, tem-se que a probabilidade da frase “*This place is great*” ser classificada como positiva é

$$P(\text{Positiva} | \text{This place is great}) = 0,000023 \times 0,6 = 0,0000138. \quad (3.11)$$

Efetuando o mesmo processo para a probabilidade da frase ser negativa, tem-se:

$$P(\text{Negativa} | \text{This place is great}) = 0,0000049 \times 0,4 = 0,00000196. \quad (3.12)$$

Portanto, tem-se que a frase “*This place is great*” é positiva, uma vez que a probabilidade de ser positiva (0,0000138) é maior que a probabilidade dessa frase ser negativa (0,00000196).

3.2 Método de VADER

O método de VADER é um dicionário e classificador de sentimentos que se baseia em regras, portanto, um método de classificação simbólico. Esse foi desenvolvido especificamente para ser utilizado em redes sociais, onde se tem um contexto vago, pouca quantidade de texto, gírias e *emoticons* (HUTTO; GILBERT, 2014).

A classificação do sentimento é feita através da separação da frase em palavras, sendo que para cada palavra da frase é atribuída uma pontuação de intensidade em uma escala de -4 (sentimento negativo) até +4 (sentimento positivo). Como por exemplo, a palavra *great* tem a intensidade de 3.1 e *horrible* -2.5. Essa pontuação é obtida através de

um dicionário que é construído utilizando o método de “*wisdom of the crowd*”, onde um grupo de pessoas é responsável por atribuir os valores de intensidade para cada palavra. Por exemplo, a frase “*This place is great*” seria classificada com

Palavra	<i>This</i>	<i>place</i>	<i>is</i>	<i>great</i>
Intensidade:				3,1

As palavras “*This*”, “*place*” e “*is*” são desconsideradas uma vez que não existem no dicionário e não expressam sentimentos. Após, essa fase inicial, utiliza-se o seguinte conjunto de regras, os quais utilizam valores empiricamente derivados para inferir a intensidade do sentimento (HUTTO; GILBERT, 2014):

- É verificado quando uma palavra que expressa sentimentos é escrita em letras maiúsculas. Neste caso, é aumentada a magnitude da intensidade do sentimento sem modificar a orientação semântica. Isso é feito somando 0,733 na intensidade do sentimento caso este tenha intensidade positiva ou subtraindo 0,733 caso este tenha uma intensidade negativa.

Palavra	<i>This</i>	<i>place</i>	<i>is</i>	<i>GREAT</i>
Intensidade:				3,1 → 3,833

- É verificado se alguma das três palavras anteriores é um advérbio intensificador. Neste caso, a intensidade do sentimento é aumentada ou diminuída em 0,293.

Palavra	<i>This</i>	<i>place</i>	<i>is</i>	<i>incredibly</i>	<i>great</i>
Intensidade:				Advérbio	3,1 → 3,393

Palavra	<i>This</i>	<i>place</i>	<i>is</i>	<i>somewhat</i>	<i>great</i>
Intensidade:				Advérbio	3,1 → 2,807

- É verificado se a frase contém a palavra “*but*”. Essa palavra indica uma troca do sentimento da frase, uma vez que o texto seguinte a palavra “*but*” expressa um sentimento dominante. Neste caso, o método multiplica a intensidade dos sentimentos expressos até a palavra “*but*” por 0,5 e os sentimentos expressos após a palavra “*but*” por 1,5.

Palavra	<i>Great</i>	<i>place</i>	<i>but</i>	<i>today</i>	<i>the</i>	<i>food</i>	<i>was</i>	<i>horrible</i>
Intensidade:	3,1 → 1,55							-2,5 → -3,75

- É verificado se a frase possui pontos de exclamação (!). Este tipo de pontuação aumenta a magnitude da intensidade sem modificar a orientação semântica. Neste caso, é adicionado um valor de 0,292 a cada ponto de exclamação, considerando um máximo de 4 pontos de exclamação.

Palavra	<i>This</i>	<i>place</i>	<i>is</i>	<i>great!</i>
Intensidade:				3,1 → 3,392

- São examinadas as três palavras anteriores, procurando a existência de uma negação que inverta a polaridade do texto. Quando é encontrada uma negação na frase, a intensidade de cada palavra de sentimento é multiplicada por -0,74.

Palavra	<i>This</i>	<i>place</i>	<i>wasn't</i>	<i>great</i>
Intensidade:			Negação	3,1 → -2,294

Após o cálculo de intensidade, é feita a normalização dessa pontuação, a partir da Equação 3.13. Neste caso, 15 é um valor fixo para aproximar o resultado final do valor máximo esperado (-1 para palavras negativas e +1 para palavras positivas).

$$\text{Pontuação Normalizada} = \frac{\text{Pontuação}}{\sqrt{\text{Pontuação}^2 + 15}} \quad (3.13)$$

$$\text{Pontuação Normalizada} = \frac{3,1}{\sqrt{3,1^2 + 15}} = 0,6249 \quad (3.14)$$

Neste caso para a frase “*This place is great*”, será atribuída uma pontuação final de 0,6249. Caso a pontuação da frase fosse menor que -1 ou maior que 1, essa seria limitada aos valores de -1 ou 1, respectivamente. Para o VADER, são consideradas frases negativas aquelas que apresentam uma pontuação de -1 até -0,5, frases neutras aquelas que apresentam uma pontuação de -0,5 até 0,5 e frases positivas aquelas que apresentam uma pontuação 0,5 até 1. Portanto, a frase “*This place is great*” com pontuação de 0,6249 seria classificada como positiva.

3.2.1 Definição de Método para Identificação de Sentimentos

Para o desenvolvimento deste trabalho optou-se pelo uso do método de VADER, uma vez que este, segundo a literatura, apresenta resultados superiores ao Método de Naive Bayes. De fato, segundo Pålsson e Szerszen (PÅLSSON; SZERSZEN, 2016), para a análise de *tweets*¹, o método de VADER apresentou uma assertividade de 72,3% enquanto o Naive Bayes apresentou uma assertividade de apenas 58,2%.

Na análise conduzida por Hutto (HUTTO; GILBERT, 2014), considerando-se a análise de *tweets*, avaliações de produtos da Amazon e editoriais do New York Times, o método de VADER apresentou uma assertividade de 96%, 63% e 55%, respectivamente, contra 56%, 49%, 44% do método de Naive Bayes. A Figura 3 apresenta um resumo dos resultados demonstrados na literatura.

¹ *Tweets* é o termo utilizado para designar as publicações realizadas na rede social Twitter.

<ul style="list-style-type: none"> ● Naive Bayes. <ul style="list-style-type: none"> ○ Análise de Tweets. <ul style="list-style-type: none"> ■ Palsson e Szerszen: 58,2% ■ Hutto e Gilbert: 56%. ○ Avaliações de Produto da Amazon. <ul style="list-style-type: none"> ■ Hutto e Gilbert: 49%. ○ Editoriais do New York Times. <ul style="list-style-type: none"> ■ Hutto e Gilbert: 44%. 	<ul style="list-style-type: none"> ● VADER. <ul style="list-style-type: none"> ○ Análise de Tweets. <ul style="list-style-type: none"> ■ Palsson e Szerszen: 72,3% ■ Hutto e Gilbert: 96%. ○ Avaliações de Produto da Amazon. <ul style="list-style-type: none"> ■ Hutto e Gilbert: 63%. ○ Editoriais do New York Times. <ul style="list-style-type: none"> ■ Hutto e Gilbert: 55%.
---	---

Figura 3 – Resultados obtidos para definição do método.

Destaca-se ainda que o método de VADER possui uma maior adaptabilidade, uma vez que esse não necessita do desenvolvimento de um *training set* específico para o tema que está sendo analisado. Desta forma, sendo o mais adequado para a análise de sentimentos na rede social Reddit, visto que os tópicos que foram selecionados neste trabalho abrangem uma quantidade diversificada de assuntos, o que inviabiliza a especialização de um *training set* para cada um dos tópicos.

3.2.2 NLTK - *Natural Language Toolkit*

A implementação de um software de análise de sentimentos pode ser feita através do desenvolvimento de um *software* por completo, ou através da utilização de *frameworks* já disponíveis. Métodos estatísticos, como o Naive Bayes, podem ser encontrados tanto em *frameworks* de aprendizado de máquina, quanto em *frameworks* voltados para o NLP. Já os métodos simbólicos, por suas implementações serem específicas para o NLP ou análise de sentimentos, somente são encontrados em *frameworks* de NLP.

O método de VADER, que será utilizado nesse trabalho, encontra-se disponível como um *package* Python (ROSSUM, 1995), ou ainda, através do *framework* NLTK (LOPER; BIRD, 2002), o qual implementa diversos métodos de NLP, incluindo métodos para análise de sentimentos. Neste trabalho, optou-se pelo uso do *framework* NLTK, devido a possibilidade de uma futura utilização de outros métodos de NLP para a análise de sentimentos.

O NLTK é um *framework open source* para Python que foi criado em 2001 na Universidade de Pensilvânia e, atualmente, é utilizado por mais de 30 universidades em diversos países (Natural Language Toolkit, 2017). Esse apresenta tanto métodos estatísticos como, MaxEnt e Naive Bayes, como métodos simbólicos, contendo mais de 50 dicionários e modelos já treinados. Além do VADER, pode-se citar o conjunto de dados *Sentiment Polarity Dataset Version 2.0*, que contém mais de 1000 filmes classificados de forma positiva e 1000 filmes classificados de forma negativa (PANG; LEE; VAITHYANATHAN,

2002). Além disso, destaca-se o *SentiWordNet* que é um dicionário com as palavras extraídas do WordNet já classificadas em positividade, negatividade e objetividade (ESULI; SEBASTIANI, 2006).

4 Criação de Base de Dados

Este capítulo tem como objetivo descrever a rede social Reddit e apresentar os tópicos que foram selecionados para a análise de sentimentos. Após, é demonstrada a ferramenta desenvolvida para a extração dos comentários destes tópicos e para a criação da base. Por fim, é apresentada a validação da ferramenta desenvolvida.

4.1 Rede Social Reddit

O *website* Reddit foi criado por Alexis Ohanian e Steve Huffman e teve seu início em 2005 como um agregador de conteúdo e, atualmente, é o vigésimo terceiro *website* mais acessado na internet e o sétimo mais acessado nos Estados Unidos da América (Alexa, 2017). Os usuários do Reddit podem enviar *links* com conteúdos externos ao Reddit ou ainda mensagens de texto. A partir desse conteúdo, os seus usuários podem votar para cima (*upvote*) ou para baixo (*downvote*), influenciando na posição do conteúdo no *website*. Além disso, seus usuários podem enviar comentários como forma de expressar sua opinião (Figura 4).

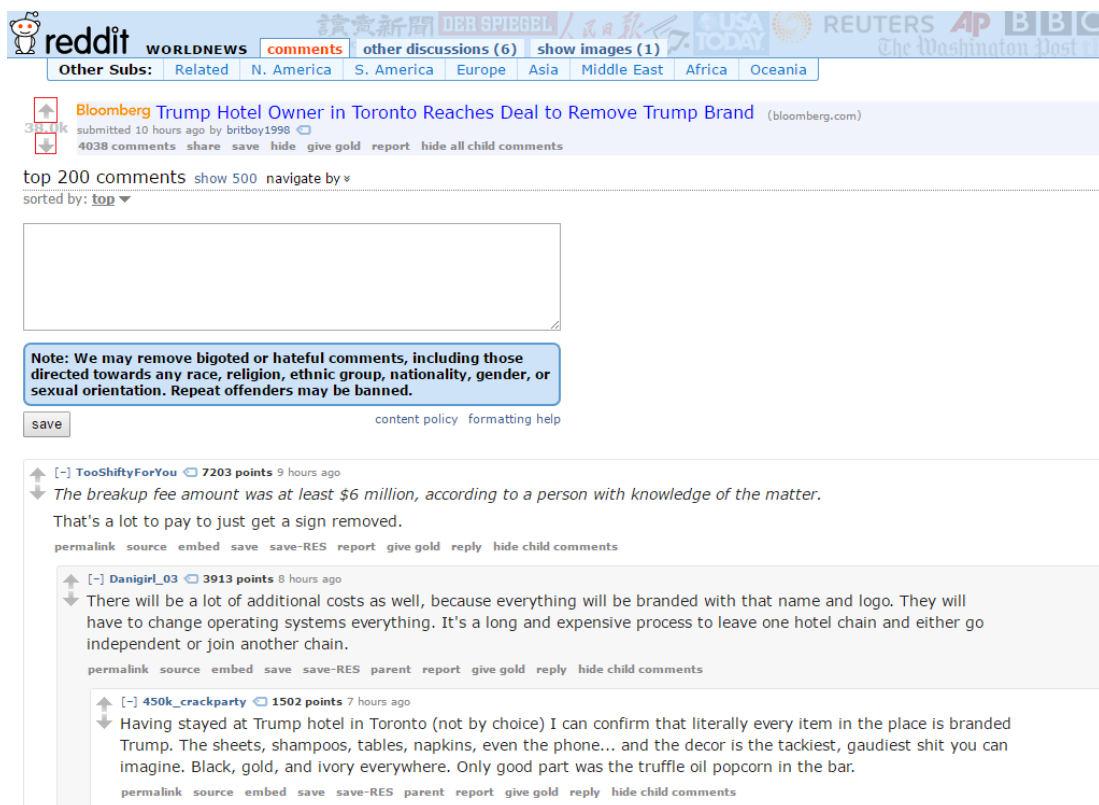


Figura 4 – *Website Reddit*: As flechas demarcadas permitem efetuarmos *upvotes* ou *downvotes*.

O conteúdo do Reddit é distribuído em *subreddits* que funcionam como comunidades. Os usuários podem se inscrever nesses *subreddits*, recebendo as atualizações na sua página inicial, sendo que dentre esses *subreddits*, destacam-se:

- */r/AskReddit*: esse *subreddit* é utilizado para fazer perguntas gerais para outros usuários do Reddit. Esse *subreddit* possui aproximadamente 16.941.540 inscritos.
- */r/worldnews*: esse *subreddit* possui as notícias de todo o mundo, possuindo aproximadamente 16.570.600 inscritos.
- */r/IAmA*: IAmA é uma estilização de 'I am a' ('Eu sou um'): a partir desse *subreddit* os usuários podem fazer perguntas ao criador de um determinado tópico. Esse *subreddit* possui aproximadamente 16.990.160 inscritos.

4.2 Extração de Dados

Para a extração de comentários do *website* Reddit e para a criação da base foi desenvolvido um *crawler* ou robô de navegação. Esse robô foi implementado na linguagem de programação Java e tem como objetivo a navegação automática no conteúdo do *website*, extraindo os dados e comentários referentes a um determinado tópico. Após, os dados são armazenados em uma base de dados criada em um banco de dados MySQL (WIDENIUS; AXMARK, 2002). Na Figura 5 tem-se a arquitetura do *software* desenvolvido. O código deste encontra-se no CD em anexo.

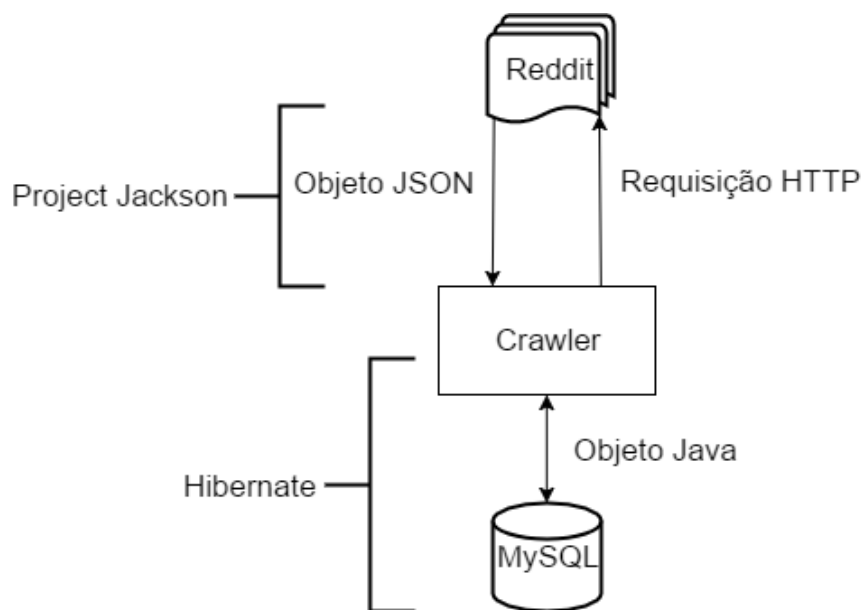


Figura 5 – Arquitetura do *Crawler*

A partir do *link* de um tópico, o robô efetua uma busca e a extração dos dados relacionados a esse tópico. Para tanto, foi utilizada a *Application Programming Interface*

(API) do Reddit, onde, inicialmente envia-se uma requisição utilizando o sufixo “.json” (Por exemplo: `<https://www.reddit.com/r/iama.json>`) e, a partir dessa requisição, o *website* retorna um objeto *JavaScript Object Notation* (JSON). Uma vez que o JSON retornado possui 68 campos e que esses não se encontram documentados, utilizou-se o *website* *jsonschema2pojo*¹ para converter o JSON retornado em um *Plain Old Java Objects* (POJO).

Após, foi utilizado o *framework* Hibernate (IVERSON, 2004) para a criação do banco de dados e para a persistência dos dados. O Hibernate é um *framework* de mapeamento objeto-relacional que tem como principal objetivo representar as tabelas de um banco de dados através de classes, ou seja, esse *framework* tem como principal característica a transformação das classes em Java em tabelas em um banco de dados relacional. No caso deste trabalho, ele é responsável pela criação das tabelas *RedditPost* e *RedditThread*, relacionadas, respectivamente, com os comentários e o tópico em questão.

4.3 Validação da Implementação

Para validação da implementação desenvolvida foi selecionado o tópico: “*Canada will welcome you, Trudeau invites refugees as Trump bans them*”. A partir desse tópico, foram extraídos todos os comentários, sendo que esses foram avaliados de forma manual. Após, esses comentários foram processados através da ferramenta NLTK, obtendo uma assertividade de 56%.

Destaca-se que uma grande parte dos comentários que foram identificados de forma incorreta, apresentavam a irônia como característica, como por exemplo, o comentário “*lol Good Luck Canada*”. A identificação da ironia através da leitura textual é difícil até mesmo para um ser-humano, uma vez que o texto pode não apresentar sinais de humor. De fato, a ironia para a análise de sentimentos é considerado um tópico problemático sendo alvo de diversos estudos (STRANISCI et al., 2016).

Observou-se que nos demais comentários que foram identificados de forma incorreta, o sentimento expresso não se encontrava no dicionário de sentimentos padrão do VADER. Desta forma, foi utilizado o método de Propagação Dupla, que possui como principal objetivo adicionar novas palavras ao dicionário padrão do VADER (QIU et al., 2011).

4.4 Método de Propagação Dupla

O método de Propagação Dupla, proposto por Qiu (QIU et al., 2011), apresenta como objetivo resolver dois problemas do NLP, que são a expansão do dicionário e também

¹ <http://www.jsonschema2pojo.org/> - O *website* *jsonschema2pojo* tem como objetivo a conversão de um esquema JSON em *Plain Old Java Objects* (*Plain Old Java Objects* (POJO)), permitindo o *download* da classe para a utilização.

a palavra “*witty*” será acrescida ao dicionário de sentimentos uma vez que essa palavra é um adjetivo que ainda não se encontra no dicionário de sentimentos.

We have a witty president.
 $\text{president} \rightarrow \text{witty}$

A segunda regra da tarefa dois consiste em verificar se uma palavra alvo possui um dependente, que por sua vez possui um adjetivo como dependente. Por exemplo, na frase “*Trump is a wholesome president.*” a palavra “*wholesome*”, que é um adjetivo, depende de “*is*” que por sua vez depende de “*president*”. Neste caso, a palavra “*wholesome*” será adicionada ao dicionário de sentimentos uma vez que essa é um adjetivo que ainda não existe no dicionário.

Trump is a wholesome president.
 $\text{president} \rightarrow \text{is} \rightarrow \text{wholesome}$

A **terceira tarefa** consiste na extração de palavras alvo a partir de palavras alvo que já se encontram na lista de alvos. A primeira regra desta tarefa verifica se a frase possui alguma palavra na lista de alvos, e verifica se essa possui alguma conjunção. Em caso positivo, a palavra após a conjunção é adicionada na lista de alvos. Por exemplo, na frase “*We have a great president and leader.*”, a palavra “*president*” que foi extraída através da primeira tarefa, possui a conjunção “*and*” que a relaciona com a palavra “*leader*”, a qual não consta na lista de alvos. Neste caso, a palavra “*leader*” será adicionada a lista de alvos.

We have a great president and leader.
 $\text{president} \rightarrow \text{and} \rightarrow \text{leader}$

A segunda regra da terceira tarefa verifica se dois substantivos possuem uma palavra dependente em comum. No caso de um desses substantivos ser uma palavra alvo, o outro também é adicionado na lista. Por exemplo, na frase “*Trump is a great president.*”, a palavra “*president*” que foi extraída através da primeira tarefa possui a palavra “*is*” como dependente, bem como a palavra “*Trump*” também possui a palavra “*is*” como dependente. Desta forma, a palavra “*Trump*” será adicionada a lista de alvos.

Trump is a great president.
 $\text{Trump} \rightarrow \text{is} \leftarrow \text{president}$

A **quarta tarefa** tem como objetivo a extração de novos sentimentos a partir de adjetivos já extraídos. A primeira regra desta tarefa, efetua a extração através das conjunções presentes no texto. Por exemplo, para a frase “*Trump is witty and clever.*”, a

palavra “*witty*” possui uma conjunção (“*and*”), que a relaciona com a palavra “*clever*”. Neste caso, a palavra “*clever*” será adicionada ao dicionário.

$$\text{Trump is } \underbrace{\text{witty and clever}}_{\text{witty} \leftarrow \text{and} \rightarrow \text{clever}}.$$

A segunda regra da quarta tarefa, consiste em verificar-se dois adjetivos dependem de uma mesma palavra. Se uma destas palavras pertence ao dicionário de sentimentos, a outra também será adicionada ao dicionário. Por exemplo, na frase “*Trump is witty, clever*”, ambas as palavras “*witty*” e “*clever*” dependem de “*Trump*”. Neste caso, será extraída a palavra “*clever*” e adicionada ao dicionário.

$$\text{Trump } \underbrace{\text{is witty, clever}}_{\text{witty} \leftarrow \text{Trump} \rightarrow \text{clever}}.$$

Por fim, é verificado se a lista de alvo ou o dicionário de sentimento sofreram alterações, ou seja, se foram adicionadas novas palavras a essas listas. Em caso positivo, as quatro tarefas serão executadas até que nenhuma palavra seja adicionada nas listas.

4.4.2 Atribuição de Pontuação às Novas Palavras

Após a execução das quatro tarefas, se faz necessário definir a pontuação relativa aos novos sentimentos e aos alvos que foram identificados através do Método de Propagação Dupla. Essa atribuição é realizada da seguinte forma, para as palavras que foram extraídas através da quarta tarefa, é utilizada a mesma pontuação da palavra relacionada a essa nova palavra. Por exemplo, para a frase “*Trump is witty and clever.*”, a palavra “*witty*” possui uma conjunção (“*and*”), que a relaciona com a palavra “*clever*”. Neste caso, a palavra “*clever*” terá a mesma pontuação da palavra “*witty*”.

Para a pontuação das palavras extraídas na tarefa dois, é realizado o seguinte processo, na extração do alvo é atribuído a este alvo a mesma pontuação do sentimento relacionado com ele. Por exemplo, na frase “*We have a great president*”, a palavra “*president*” será adicionada a lista de alvos de sentimentos com a mesma pontuação de “*great*”, a qual está relacionada. Após, essa pontuação é atribuída a palavra extraída a partir da tarefa dois. Por exemplo, no comentário “*Trump is a wholesome president.*”, na segunda tarefa, a palavra “*wholesome*”, que foi extraída a partir da palavra “*president*”, irá receber a mesma pontuação da palavra “*great*”.

4.4.3 Textos utilizados para o método de Propagação Dupla

Para utilização do método de Propagação Dupla, é necessário utilizar-se de um conjunto de textos para que sejam executadas as quatro tarefas. Segundo Qiu (QIU et al.,

2011), como as palavras tem diferentes significados em diferentes contextos, é recomendado que este conjunto de textos pertença ao mesmo contexto que está sendo efetuada a análise de sentimentos. Por exemplo, ao utilizar um conjunto de dados sobre notícias gerais, pode-se determinar que a palavra “*gucci*” é uma gíria similar a palavra “*good*”. Porém, o uso dessa palavra em comentários de roupas apresenta um outro contexto, se referindo a uma marca de roupas.

A fim de se determinar o melhor conjunto de textos para utilização do método de Propagação Dupla, foram utilizados dois conjuntos de dados de testes. No primeiro conjunto de textos, considerou-se os comentários de todos os usuários que comentaram o tópico “*Canada will welcome you, Trudeau invites refugees as Trump bans them*”. Mais especificamente, foram considerados os últimos 1000 comentários destes usuários, totalizando 780063 comentários.

No segundo conjunto de textos, foram considerados comentários de usuários em tópicos que apresentam o mesmo tema do tópico “*Canada will welcome you, Trudeau invites refugees as Trump bans them*”. Nesta avaliação, foram utilizados os seguintes tópicos:

- *Donald Trump to strip all funding from State Dept team promoting women’s rights around the world - Leaked plan comes as First Daughter Ivanka defends her father’s record with women*: esse tópico contém 9246 comentários e encontra-se disponível em <https://www.reddit.com/r/worldnews/comments/67ivae/donald_trump_to_strip_all_funding_from_state_dept/>. Esse tópico refere-se a decisão do presidente dos Estados Unidos da América, Donald Trump, em remover os fundos de promoção ao direito das mulheres.
- *Sweden asks the U.S. to explain Trump comment on Sweden*: esse tópico contém 10927 comentários e encontra-se disponível em <https://www.reddit.com/r/worldnews/comments/5uzetf/sweden_asks_the_us_to_explain_trump_comment_on/> e se refere aos comentários feitos do presidente dos Estados Unidos da América, Donald Trump, sobre a Suécia.
- “*Canada will welcome you, Trudeau invites refugees as Trump bans them*”: esse tópico contém 9113 comentários e encontra-se disponível em <https://www.reddit.com/r/worldnews/comments/5qqa51/canada_will_welcome_you_trudeau_invites_refugees/> e refere-se a declaração do primeiro ministro canadense sobre decisão de receber refugiados. Nesta declaração, o primeiro ministro canadense afirma que os refugiados serão bem-vindos no Canadá.

Após a utilização do Método de Propagação Dupla utilizando esses dois conjuntos de textos, foram obtidos os seguintes resultados para o tópico “*Canada will welcome you, Trudeau invites refugees as Trump bans them*”:

- 58% de assertividade na utilização do dicionário padrão.
- 59% de assertividade na utilização do dicionário padrão com palavras extraídas a partir dos últimos 1000 comentários de todos os usuários que comentaram o tópico em questão.
- 62% de assertividade na utilização do dicionário padrão com palavras extraídas a partir de comentários postados em tópicos relacionados ao mesmo assunto.

Observa-se que os melhores resultados foram obtidos utilizando-se o método de Propagação Dupla utilizando textos com assuntos relacionados. Desta forma, essa será a abordagem utilizada no restante do trabalho.

4.5 Restrição de Palavras Alvos

Verificou-se que após a utilização do Método de Propagação Dupla, que os comentários que não foram identificados de forma correta são comentários como: *"...These laws are not "racist", morons keep hysterically throwing that word around and its losing all meaning..."*, aonde que o sentimento expresso é negativo, porém, a pessoa expressa um sentimento a favor da notícia em questão.

Para a solução deste problema, foram extraídos somente os sentimentos relacionados a uma determinada palavra alvo. Por exemplo, no tópico *"Canada will welcome you," Trudeau invites refugees as Trump bans them*", é possível que a mesma pessoa expresse uma opinião positiva com relação ao *Canadá* e negativa com relação a *Trudeau*. Ao se determinar que desejamos analisar comentários sobre a palavra *"Trudeau"*, pode-se definir *Trudeau* como a palavra alvo para a análise, sendo consideradas somente as frases relacionadas com essa palavra. Por exemplo, *"Come enjoy our 15% tax :D edit: Trudeau is the most fake person ever. He always pulls this shit but undercuts us Canadians all the time. He ain't getting voted in next election."*, será analisada somente a frase dependente da palavra *"Trudeau"*, que será *"... is the most fake person ever"*, evitando que a primeira frase *"Come enjoy our 15% tax :D..."*, que representa um sentimento positivo, afete a análise.

Restringindo a análise a palavra alvo *"Trump"* para extração dos comentários do tópico *"Canada will welcome you," Trudeau invites refugees as Trump bans them*", foi obtida a assertividade de 65%. A partir destes resultados, optou-se pela utilização do método de VADER, combinado com o método de Propagação Dupla e a restrição de alvos para a execução dos testes que serão apresentados no Capítulo 5.

5 Resultados Obtidos

Para realização dos testes, foram selecionados 6 tópicos distribuídos em duas categorias diferentes, que são: comentários políticos e comentários sobre filmes. Optou-se pela escolha de duas categorias distintas com o objetivo de avaliar a assertividade do método desenvolvido em diferentes contextos. Como a validação dos resultados é efetuada de forma manual, e devido a restrições de tempo no desenvolvimento deste trabalho, não foi possível efetuar a análise sobre mais tópicos ou categorias.

5.1 Comentários Políticos

De forma a realizar testes sobre um contexto de comentários políticos, foram executados testes sobre os comentários listados abaixo, e que foram descritos Seção 4.4.3, sendo utilizado o Método de Propagação Dupla e a restrição de palavra alvo “*Trump*”. Para a aplicação do Método de Propagação Dupla, utilizou-se como conjunto de textos os próprios comentários destes tópicos.

- *Donald Trump to strip all funding from State Dept team promoting women’s rights around the world - Leaked plan comes as First Daughter Ivanka defends her father’s record with women.*
- *Sweden asks the U.S. to explain Trump comment on Sweden.*
- *“Canada will welcome you,” Trudeau invites refugees as Trump bans them.*

Na Figura 6, tem-se os resultados obtidos. Observa-se que a assertividade foi próxima aos valores obtidos por Hutto e Gilbert (HUTTO; GILBERT, 2014) para avaliações de produtos da Amazon e para editoriais do New York Times. Porém, os valores foram inferiores aos obtidos na avaliação de *tweets*.

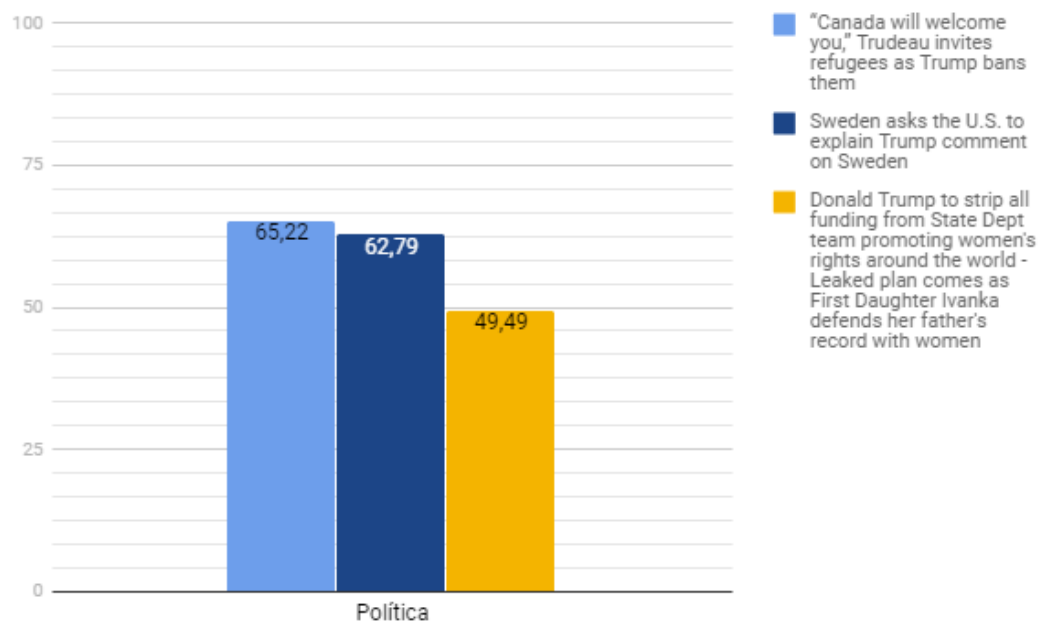


Figura 6 – Assertividade da ferramenta desenvolvida com relação a comentários de política.

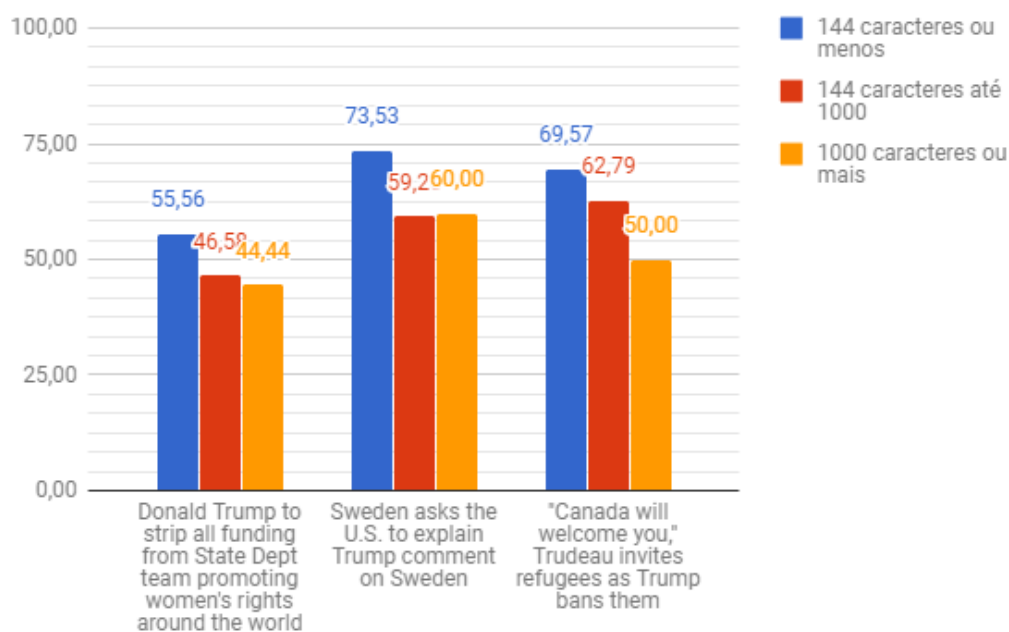


Figura 7 – Assertividade da ferramenta desenvolvida com relação a comentários de política por tamanho.

De forma a avaliar se os resultados foram ocasionados pela diferença de contexto ou por seus corpos de texto terem diferentes tamanhos, foi também conduzida uma análise separando todos os comentários de política em três categorias, comentários com 144 caracteres ou menos, similares a “*tweets*”, tópicos com menos de 1000 caracteres e tópicos com 1000 caracteres ou mais. Na Figura 7 tem-se os resultados obtidos através da análise através da quantidade de caracteres. Observa-se que o método apresenta uma diminuição na sua assertividade com o aumento do texto. Essa tendência é observada nos trabalhos de Hutto e Gilbert (HUTTO; GILBERT, 2014) onde “*tweets*” analisados apresentaram maior assertividade do que editoriais do *New York Times* e avaliações de produto da *Amazon*.

5.2 Comentários sobre Filmes

Os mesmos testes foram realizados utilizando a palavra “*movie*” como alvo, em tópicos de cinema. O método VADER, foi utilizado usando o método de Propagação Dupla, sendo que os textos selecionados para a expansão e, para a posterior análise foram obtidos dos comentários dos tópicos:

- *Official Discussion - mother! [SPOILERS]*: esse tópico contém 5297 comentários e encontra-se disponível em <https://www.reddit.com/r/movies/comments/706y1p/official_discussion_mother_spoilers/> e apresenta a avaliação do filme “*Mother!*”.
- *Official Discussion: Gerald’s Game [SPOILERS]*: esse tópico contém 892 comentários e encontra-se disponível em <https://www.reddit.com/r/movies/comments/73g2fx/official_discussion_geralds_game_spoilers/> e apresenta a avaliação do filme “*Gerald’s Game*”.
- *Official Discussion: The Mummy (2017) [SPOILERS]*: esse tópico contém 1333 comentários e encontra-se disponível em <https://www.reddit.com/r/movies/comments/6g5lmo/official_discussion_the_mummy_2017_spoilers/> e apresenta a avaliação do filme “*The Mummy*”.

Na Figura 8 tem-se os resultados obtidos. Observa-se que a assertividade também se mostra próxima aos valores obtidos por Hutto e Gilbert (HUTTO; GILBERT, 2014). Observa-se neste caso que os resultados foram superiores aos apresentados na Seção 5.1. Essa melhora na assertividade se deve ao fatos dos tópicos de filme encorajarem ao leitor a efetuar a avaliação do filme em questão, enquanto os comentários sobre política apresentam no conteúdo em um *link* direto para a notícia original.

Posteriormente, foi conduzida uma análise separando os comentários de filmes, de acordo com o tamanho do texto, sendo considerados: comentários com 144 caracteres ou menos (similares a “*tweets*”), comentários com menos de 1000 caracteres e comentários

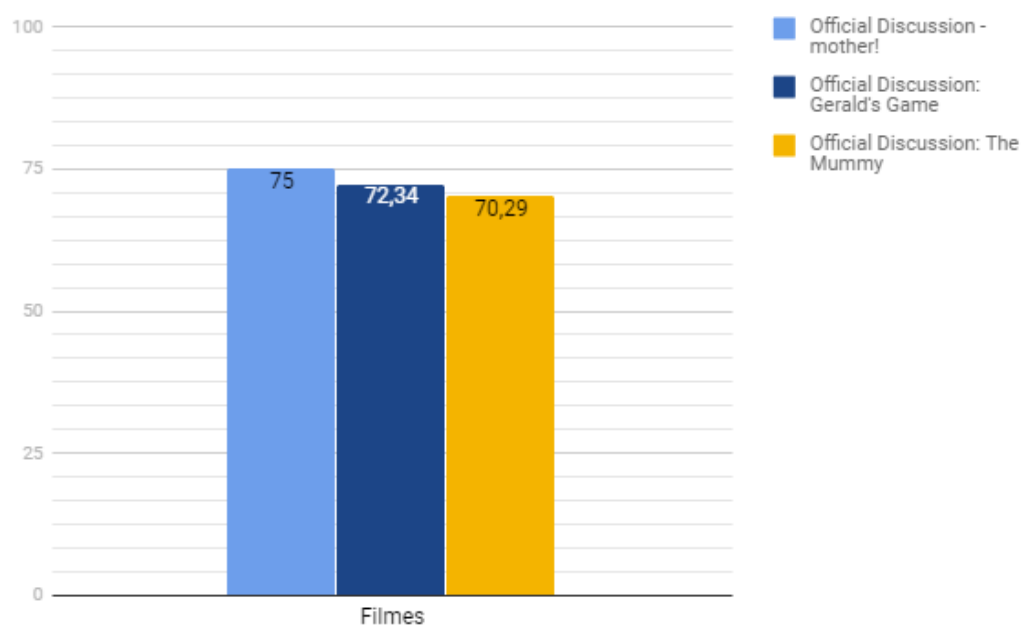


Figura 8 – Assertividade da ferramenta desenvolvida com relação a comentários de filmes.

com 1000 caracteres ou mais. Na Figura 9 tem-se os resultados obtidos através da análise através da quantidade de caracteres. Observa-se que neste caso, também houve uma queda na assertividade com o aumento do texto.

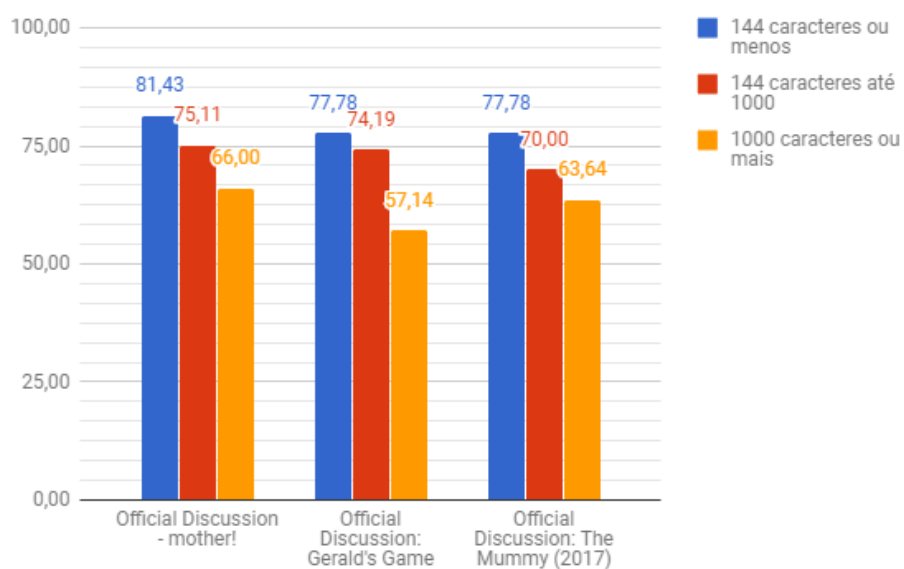


Figura 9 – Assertividade da ferramenta desenvolvida com relação a comentários de filmes por tamanho.

6 Conclusão Final

Através deste trabalho, observou-se que os métodos mais utilizados para análise de sentimentos são o Método de Naive Bayes (estatístico) e o Método de VADER (simbólico). Dentre estes, optou-se pela utilização do método de VADER, uma vez que de acordo com a literatura, esse apresenta um desempenho superior ao método de Naive Bayes, na análise de sentimentos nas avaliações de produtos da Amazon, editoriais do New York Times e mais importante, na análise de *Tweets* da rede social Twitter (PÅLSSON; SZERSZEN, 2016).

A partir dos resultados obtidos verifica-se que a utilização do método de VADER isolado apresenta resultados insatisfatórios. Desta forma, foi utilizado o método de Propagação Dupla e combinado com a escolha de palavras alvo para a análise. Neste caso, os resultados apresentaram melhorias de mais 7%. De fato, no caso do Método de VADER isolado, obteve-se uma assertividade de 58%. Já utilizando o Método de Propagação Dupla com palavras alvos, obteve-se uma assertividade de 65%.

Foram analisados 6 tópicos divididos em duas categorias, comentários políticos e comentários de filmes. Estes tópicos foram analisados por completo e posteriormente, considerando o número de caracteres. Os resultados obtidos demonstraram que a aplicação do método desenvolvido sobre comentários de filmes apresenta uma maior assertividade do que a aplicação do método sobre comentários políticos. Isso se deve ao fato de que nos tópicos relacionados com filmes, é pedido a opinião ou avaliação dos usuários sobre aquele filme, aumentando a utilização de expressões simples como “... *este filme foi ruim*...”. Já a partir da análise de assertividade por quantidade de caracter, foi possível concluir que o VADER tende a perder assertividade com o aumento no número de caracteres em um comentário. De fato, quanto maior o número de caracteres, menor a assertividade do método. Essas tendências já haviam sido observadas nos trabalhos de Hutto e Gilbert (HUTTO; GILBERT, 2014).

6.1 Trabalhos Futuros

Como trabalhos futuros, segere-se a análise de comentários do Reddit através do “*Naive Bayes*” de forma a comparar com o método de VADER na sua utilização no Reddit.

Sugere-se ainda, a incorporação do histórico do usuário na análise de sentimentos. Ou seja, para definirmos se um usuário está sendo irônico, por exemplo, seria interessante verificar se este sempre elogiou determinado produto, e somente neste determinado comentário ele está criticando este.

Referências

- Alexa. *Alexa Top 500 Global Sites*. 2017. <Disponível em: <<http://www.alexa.com/topsites/>>>. Acesso em: 27 de Fevereiro de 2017. Citado 2 vezes nas páginas 17 e 35.
- Apache OpenNLP. *Apache OpenNLP*. 2017. <Disponível em: <<https://opennlp.apache.org/>>>. Acesso em: 27 de Fevereiro de 2017. Citado na página 17.
- BERGER, A. L.; PIETRA, V. J. D.; PIETRA, S. A. D. A maximum entropy approach to natural language processing. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 22, n. 1, p. 39–71, mar. 1996. ISSN 0891-2017. Disponível em: <<http://dl.acm.org/citation.cfm?id=234285.234289>>. Citado na página 27.
- BRILL, E. A simple rule-based part of speech tagger. In: *Proceedings of the Third Conference on Applied Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. (ANLC '92), p. 152–155. Disponível em: <<http://dx.doi.org/10.3115/974499.974526>>. Citado na página 20.
- ESULI, A.; SEBASTIANI, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In: *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*. [s.n.], 2006. p. 417–422. Disponível em: <SENTIWORDNET:Apubliclyavailablelexicalresourceforopinionmining>. Citado na página 34.
- HANCOX, P. J. *A brief history of Natural Language Processing*. 2017. <Disponível em: <http://www.cs.bham.ac.uk/~pjh/sem1a5/pt1/pt1_history.html/>>. Acesso em: 02 de Abril de 2017. Citado na página 19.
- HEARST, M. A. Support vector machines. *IEEE Intelligent Systems*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 13, n. 4, p. 18–28, jul. 1998. ISSN 1541-1672. Disponível em: <<http://dx.doi.org/10.1109/5254.708428>>. Citado na página 27.
- HUTTO, C. J.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: ADAR, E. et al. (Ed.). *ICWSM*. The AAAI Press, 2014. ISBN 978-1-57735-659-2. Disponível em: <<http://dblp.uni-trier.de/db/conf/icwsml/icwsml2014.html#HuttoG14>>. Citado 6 vezes nas páginas 30, 31, 32, 43, 45 e 47.
- IVERSON, W. *Hibernate: A J2EE(TM) Developer's Guide*. [S.l.]: Addison-Wesley Professional, 2004. ISBN 0321268199. Citado na página 37.
- JACKSON, P.; MOULINIER, I. *Natural Language Processing for Online Applications: Text retrieval, extraction and categorization*. [S.l.]: John Benjamins Publishing Company, 2007. Citado na página 17.
- LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ETMTNLP '02), p. 63–70. Disponível em: <<http://dx.doi.org/10.3115/1118108.1118117>>. Citado na página 33.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715. Citado na página 29.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. [S.l.]: MIT Press, 1999. Citado 5 vezes nas páginas 19, 20, 22, 23 e 28.

Natural Language Toolkit. *Natural Language Toolkit*. 2017. <Disponível em: <<http://www.nltk.org/>>>. Acesso em: 27 de Fevereiro de 2017. Citado 2 vezes nas páginas 17 e 33.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (EMNLP '02), p. 79–86. Disponível em: <<https://doi.org/10.3115/1118693.1118704>>. Citado 2 vezes nas páginas 27 e 34.

PÅLSSON, A.; SZERSZEN, D. *Sentiment Classification in Social Media: An Analysis of Methods and the Impact of Emoticon Removal (Dissertation)*. 2016. Citado 2 vezes nas páginas 32 e 47.

QIU, G. et al. Opinion word expansion and target extraction through double propagation. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 37, n. 1, p. 9–27, mar. 2011. ISSN 0891-2017. Disponível em: <http://dx.doi.org/10.1162/coli_a_00034>. Citado 2 vezes nas páginas 37 e 41.

ROSSUM, G. *Python Reference Manual*. Amsterdam, The Netherlands, The Netherlands, 1995. Citado na página 33.

SHANNON, C. E.; WEAVER, W. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, p. 379–423, 623–656, July, October 1948. Citado na página 20.

Spacy. *Spacy*. 2017. <Disponível em: <<https://spacy.io/>>>. Acesso em: 27 de Fevereiro de 2017. Citado na página 17.

Stanford CoreNLP. *Stanford CoreNLP*. 2017. <Disponível em: <<http://stanfordnlp.github.io/CoreNLP/>>>. Acesso em: 27 de Fevereiro de 2017. Citado na página 17.

STRANISCI, M. et al. Annotating sentiment and irony in the online italian political debate on #labuonascuola. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. [s.n.], 2016. Disponível em: <<http://www.lrec-conf.org/proceedings/lrec2016/summaries/1063.html>>. Citado na página 37.

WIDENIUS, M.; AXMARK, D. *Mysql Reference Manual*. 1st. ed. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2002. ISBN 0596002653. Citado na página 36.