

Sistema de Monitoração e Análise de Comentários nas Mídias Sociais

Jader Fabiano Batista Marques, Fernanda dos Santos Cunha, Anita M^a da Rocha Fernandes

Curso de Ciência da Computação
Centro de Ciências Tecnológicas da Terra e do Mar – CTTMAR
Universidade do Vale do Itajaí – Campus KobraSol – São José/SC – Brasil

jaderfbmg@gmail.com, {fernanda.cunha, anita.fernandes}@univali.br

Abstract. *The large number of people using social networks and their use for display of feelings led several companies concerned with identifying what your potential customers are expressing. The sheer volume of data makes it impractical for a social media analyst measure and understand the feelings of their customers. This paper presents the development of a tool for monitoring company profiles on Twitter, to identify the feeling of every comment and their level of impact for the company. The tool uses natural language processing techniques and sentiment analysis and achieved an accuracy of 85.1 % in identifying the positive or negative nature of the comments.*

Resumo. *O grande número de pessoas utilizando as redes sociais e o seu uso para exposição de sentimentos levaram várias empresas a preocupação de identificar o que seus potenciais consumidores estão expressando. O enorme volume de dados torna impraticável para um analista de mídia social mensurar e compreender os sentimentos de seus clientes. Este artigo apresenta o desenvolvimento de uma ferramenta para monitorar perfis de empresas no Twitter, a fim de identificar o sentimento de cada comentário e seu nível de impacto para a empresa. A ferramenta utiliza técnicas de processamento de linguagem natural e análise de sentimento e alcançou uma precisão de 85,1% na identificação da natureza positiva ou negativa dos comentários.*

1. Introdução

Nos últimos anos, o boom das mídias sociais influenciou diretamente em inúmeras empresas a se renderem ao mundo digital. O fato é que, diariamente, novos perfis corporativos são criados em diversas redes sociais, porém poucas são as empresas que realmente entendem a importância e a dinâmica das redes sociais [Oliveira 2013].

A velocidade de propagação das informações geradas nestas redes se dá nos dois casos, tanto nos negativos, quanto nos positivos. Com isso, outra oportunidade que surgiu foi a possibilidade de monitorar as conversas dos internautas, a fim de antecipar tendências e avaliar como está seu nicho de mercado. Medir o que o consumidor fala sobre a concorrência, fundamentando a sobrevivência a um mercado com clientes cada vez menos fiéis e cada vez mais ávidos por empresas transparentes e socialmente responsáveis [Grigoletti 2011].

Monitorar as redes sociais, pensando em estratégias e executando ações para divulgar produtos e serviços, pesquisar público-alvo e novas vertentes para o negócio e, claro, responder a questões específicas de internautas, especialmente quando são críticas à empresa. Essas são as funções básicas de um analista de mídias sociais, profissional cada vez mais requisitado por companhias de todas as áreas de atuação, que apostam na velocidade da web e na expansão das redes sociais para alavancar seus ganhos [Kopschitz 2011].

Os consumidores preferem fazer reclamações em redes sociais antes mesmo de entrar em contato com o serviço de atendimento ao cliente, sobrecarregando o analista de mídias sociais e, com isso, não atendendo aos consumidores em tempo hábil. Para [Grigoletti 2011], um dos aspectos que mudou no mercado consumidor com as mídias sociais foi o poder de voz dos consumidores e, conseqüentemente, a visibilidade das reclamações sobre produtos e serviços. São usuários que expõem seus problemas e criticam a marca que não atenderam suas expectativas. Por outro lado, empresas com boa visão estratégica aproveitam a oportunidade para dialogar, resolvendo problemas, criando vínculos e transformando possíveis crises de imagem em um fortalecimento institucional.

Contudo, existem técnicas que podem ser aplicadas em cima do conteúdo gerado dentro das mídias sociais. Análise de sentimentos é uma das técnicas que podem ser empregadas nas informações produzidas nas mídias sociais. Segundo [Rodrigues et al 2010], a Análise de Sentimentos visa identificar o sentimento que os usuários apresentam a respeito de alguma entidade de interesse (um produto específico, uma empresa, um lugar, uma pessoa, dentre outros). Já para [Gomes 2012], a Análise de Sentimentos é o estudo computacional de opiniões, sentimentos e emoções expressadas no texto.

Conforme [Liu 2010], a informação textual pode ser classificada em dois tipos principais: fatos e opiniões. Os fatos são expressões objetivas sobre entidades, eventos e suas propriedades, já opiniões são geralmente expressões subjetivas que descrevem os sentimentos e avaliações das pessoas sobre determinada entidade. Palavras de opinião são palavras que são comumente usadas para expressar sentimentos negativos e positivos. Por exemplo, “belo, bom, maravilhoso e incrível” são palavras de opiniões positivas, e “ruim, pobre e terrível” são palavras de opiniões negativas. Frases e palavras de opiniões são fundamentais para análise de sentimentos por razões óbvias.

Com isso, pode-se conceituar que o processamento de linguagem natural é a compreensão de uma informação falada ou escrita por meio de regras equivalentes àqueles existentes na comunicação linguística humana. A fim de melhor compreender como ocorre o processamento de uma informação textual, deve-se analisar as fases do processamento de linguagem natural, que são: Análise Morfológica, Análise Sintática, Análise Léxica, Análise Semântica e Análise Pragmática [Cesário Júnior 2010].

Sendo assim, foi desenvolvida uma ferramenta para monitoração e análise de comentários no Twitter, a fim de auxiliar as empresas a melhor administrarem o relacionamento com seu público-alvo. Este artigo apresenta nas próximas seções a teoria envolvida, alguns trabalhos correlatos, o desenvolvimento do projeto e seus resultados e conclusões.

2. Fundamentação Teórica

A informação textual no mundo pode ser amplamente categorizada em dois principais tipos: fatos e opiniões. Fatos são expressões objetivas sobre entidades, eventos e suas propriedades.

Opiniões são geralmente expressões subjetivas que descrevem sentimentos, avaliações ou sensações de pessoas sobre entidades, eventos e suas propriedades [Liu 2010].

A análise de sentimento é o campo de estudo que analisa opiniões, sentimentos, avaliações, atitudes e emoções das pessoas relativas a entidades, tais como, produtos, organizações, serviços, indivíduos, eventos e seus atributos [Liu 2012].

Um sentimento pode ser visto como valores pessoais (positivos, negativos ou neutros) expressos sobre uma determinada entidade. Porém, um sentimento normalmente não é expresso de forma clara e precisa. A Análise de Sentimentos basicamente tenta inferir os sentimentos de pessoas baseado nas expressões de linguagem [Koblitz 2010].

Existem duas técnicas principais para se realizar a análise de sentimento, uma usando classificadores de aprendizagem de máquina e outra usando lexicons, um lexicon é um conjunto de palavras e expressões associadas a algumas definições. Sua estrutura lembra um dicionário, onde uma palavra está associada a sua definição.

No lexicon as palavras e expressões podem estar associadas a informações de polaridade, classe gramatical, etc. Em uma abordagem que usa um lexicon, um documento é classificado de acordo com as palavras e expressões do documento avaliado que possuem informações no lexicon [Pang et al 2002].

Classificadores de aprendizagem de máquina são classificadores capazes de adquirir conhecimento utilizando técnicas computacionais, eles realizam classificações baseadas em experiências acumuladas anteriormente por meio de treinamento. Para a análise de sentimento, essa abordagem envolve a construção de classificadores a partir de instâncias, textos ou sentenças rotuladas para classificar novas entradas [Pang et al 2002].

Processamento de Linguagem Natural é um subcampo da Inteligência Artificial (PLN) que estuda os problemas de geração automatizada e compreensão de linguagens naturais humanas [Correia 2011]. Segundo [Chowdhury 2003], PLN é uma área de pesquisa e aplicação que explora como computadores podem ser usados para entender e manipular texto ou discurso para fazer coisas úteis. Já para [Garcia 2010] o processamento de linguagem natural engloba um vasto conjunto de tarefas que vão desde tratamentos superficiais do texto até análises profundas do ponto de vista sintático e semântico.

3. Trabalhos correlatos

Segundo [Pang e Lee 2008], descobrir como as pessoas pensam sempre foi objeto de interesse de grandes empresas, instituições governamentais entre outras, e com a popularização das mídias sociais, em especial o Twitter que é uma plataforma que fornece acesso à uma enorme massa de dados subjetivos, a tarefa de identificar o sentimento classificando qual emoção um *tweet* possui passou a ser foco de diversas pesquisas. Inúmeras aplicações estão sendo implementadas nesta área, desde ferramentas que apenas avaliam a polaridade dos sentimentos até aplicações que tentam prever tendências de

mercado, predições de acontecimentos, bem como na política para mensurar pesquisas sobre eleições e/ou identificar o sentimento da população sobre a administração de um determinado governo.

No estudo realizado por [Bollen et al 2009] com objetivo de analisar a flutuação de sentimento das mensagens e buscar eventos sócio-econômicos e políticos que pudessem estar relacionados às oscilações identificadas, os autores fizeram o uso da ferramenta POMS-ex (Profile of Mood States versão estendida) para classificar *tweets* em 6 categorias de humor: tensão, depressão, raiva, vigor, cansaço e confusão. Os resultados foram comparados com flutuações do mercado de ações índices de preços do petróleo e eventos populares, tais como eleições presidenciais. Eles descobriram que eventos na esfera social, política, cultural e econômica tem significativo efeito na dimensão do humor da população.

Uma importante característica do Twitter é que por ser em tempo real, quando algum evento importante ocorre (como um terremoto) muitas pessoas publicam vários *posts* no Twitter e isso pode ser amplamente utilizado. No trabalho de [Sakaki et al 2010] as interações em tempo real desses eventos foram investigadas, propondo-se um algoritmo para monitorar o Twitter e detectar o evento alvo, através de um modelo probabilístico que pode encontrar o centro e a trajetória do terremoto. O sistema desenvolvido detecta instantaneamente o terremoto e envia e-mails para usuários registrado. As notificações são entregues de forma ágil e mais rápida do que os anúncios oficiais pelo rádio e televisão.

Para auxiliar empresas a tomarem decisões em suas campanhas de marketing [Li e Li 2011] acompanharam a opinião dos clientes em relação a serviços e produtos para permitir que as empresas tomassem decisões com mais agilidade. Para reduzir o esforço necessário de extrair a emoção expressa numa publicação eles fizeram uso dos emoticons presentes no *tweet*, já que cerca de 87% dos *tweets* coletados possuíam o sentimento representado pelo emoticon. No entanto, a utilização desse método descarta uma grande quantidade de *tweets* que não contém emoticons em suas publicações, limitando o uso da ferramenta.

Dentro dos trabalhos citados, verificou-se que nenhum deles buscou analisar o sentimento de usuários de companhias aéreas e também não ocorreu a classificação por níveis de criticidade alertando a empresa qual post é o mais crítico em uma determinada situação. Este artigo preenche essa lacuna e apresenta uma ferramenta de monitoramento em tempo real dos comentários e sua criticidade para empresas dessa área de atuação.

4. A Ferramenta Desenvolvida

A ferramenta foi desenvolvida em dois módulos distintos e uma interface para apresentação dos dados e atendimento do analista de mídia social (Figura 1).

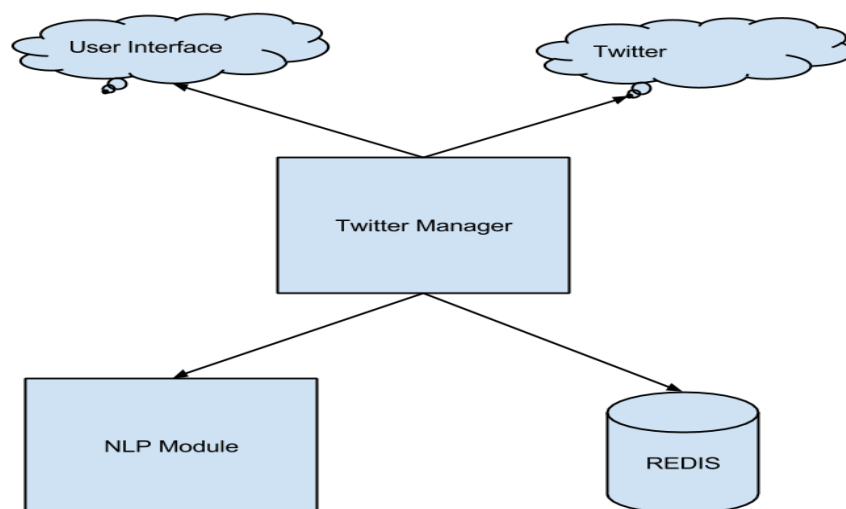


Figura 1. Visão geral da ferramenta.

A Figura 1 representa a arquitetura do sistema. O Redis é o banco de dados onde são persistidos os posts do Twitter e suas informações complementares, Redis é um banco de dados em memória não relacional (NoSQL) com persistência em disco de alta disponibilidade. TreeTagger é o "NLP Module" responsável por realizar o processamento da linguagem natural, tem como objetivo efetuar a análise morfo-sintática definindo os sintagmas de cada palavra contida em um tweet. O módulo Twitter Manager é responsável pela comunicação com a API do Twitter, gerencia o fluxo de dados dentro do sistema e persiste esses dados dentro do Redis. A API do Twitter é responsável de possibilitar a autenticação dos usuários no sistema vinculados com uma conta no Twitter, após os usuários estarem autenticados a API fornece métodos para busca de tweets juntamente com todas as informações referentes a cada tweet, por exemplo: nome do usuário, data de nascimento, identificador, cidade, país, etc.

A Interface do usuário ("User Interface") permite que o analista visualize todos os comentários postados no Twitter referente ao perfil da empresa. A apresentação se dá por meio de duas listas, uma contendo os comentários positivos e outra com os comentários identificados como negativos, ambas ordenadas pelo nível de criticidade. Por esta interface permite-se ao analista efetuar toda a interação com os responsáveis por cada post no Twitter.

Após cada novo comentário ser capturado do Twitter realiza-se sua análise morfo-sintática colocando os sintagmas referentes a cada palavra, e após realiza-se a análise de sentimento e a análise de criticidade do comentário dentro da área de domínio (Figura 2).

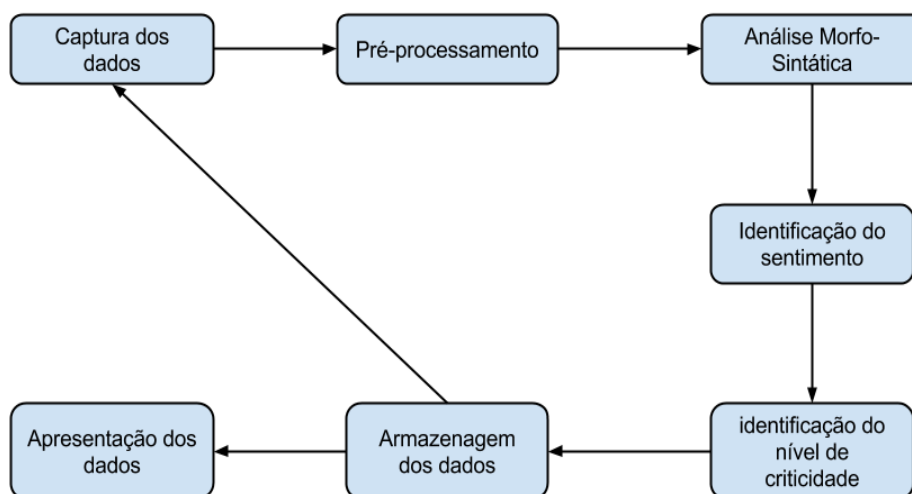


Figura 2. Fluxo análise comentário.

A captura dos dados é realizada diretamente no Twitter através do uso da REST API, para isso, fez se necessário à criação de uma aplicação na página de desenvolvimento da mídia social para adquirir a chave de acesso e uma chave secreta da aplicação.

Na etapa de pré-processamento foi convertido os caracteres da sentença para caixa baixa, remoção de caracteres a partir de três ocorrências sucessivas e substituídos por apenas um equivalente, após isso foi realizado a remoção de links e imagens, identificação de HashTags e emoticons.

Para a análise morfo-sintática foi utilizado a ferramenta TreeTagger para anotação de texto, onde essa ferramenta define os sintagmas de cada palavra em uma sentença, o TreeTagger é um etiquetador part of speech (POS). Utiliza um modelo probabilístico baseado em árvores binarias de decisões.

A realização da análise dos comentários, primeiramente foi necessário realizar uma análise visual para identificar e extrair padrões e formas de expressar o sentimento na mídia. Como o Twitter limita a inserção de cada comentário a 140 caracteres, com isso percebeu se que os usuários são mais objetivos na forma de expressar seus sentimentos. Com isso, foi feito um levantamento de adjetivos que tornam a frase tanto negativa como positiva, dentre os positivos alguns como: “ótimo, maravilhoso, excelente” e negativo “ruim, péssimo, horrível”.

Também foram extraídos verbos de caráter positivo e negativo, dentre eles: “gostar, adorar, amar, estragar, odiar, detestar”. Após isso, foram abordadas as expressões que dão sentimento a uma sentença, por exemplo: “muito bom, muito melhor, haja paciência, falta de respeito”.

Para a definição da polaridade de uma sentença é atribuído um peso a cada etapa do processo. A cada palavra de natureza positiva encontrada na sentença é acrescido um ponto, da mesma forma para cada palavra de natureza negativa é decrescido um ponto. Já para cada expressão positiva é somado dois pontos e para cada expressão negativa é diminuído dois pontos.

O resultado dessa etapa é multiplicado por 1 ou -1, caso a frase não contenha algum adjetivo ou expressão é atribuído 1, caso contrário é atribuído -1 a multiplicação.

Se a frase contiver algum emoticon é somado ou diminuído cinco pontos de acordo com a natureza deste emoticon.

Com o resultado obtido é definido a polaridade da seguinte forma: se o resultado for menor ou igual a zero, essa sentença é de natureza negativa, se a o resultado for maior que zero a sentença é de natureza positiva. Essa ferramenta não possui lista de comentários que não expressam opinião, sendo assim, todos os comentários que são de natureza neutra são colocados na lista de comentários negativos, cabendo ao analista avaliar em qual melhor lista rotulá-lo e definir qual sua ordem de atendimento.

A análise de criticidade ocorre após a identificação do sentimento da sentença. Sendo essa sentença de cunho negativo é verificado o grau de impacto do comentário com base no cadastro pré-definido pela empresa. Neste cadastro estão os dados relativos a todos os graus de impacto definidos pelo analista, a saber: número *retweets*, número de *likes*, número de seguidores e algumas palavras chaves. Essas informações são capturadas juntamente com o comentário diretamente do perfil do Twitter.

Para o analista de mídia social ter acesso aos comentários é necessário que a empresa de permissão para ferramenta acessar o perfil da empresa no Twitter. Após isso a ferramenta inicia a captura dos comentários e os rotula de acordo com sua natureza.

A tela de monitoramento possui duas listas de comentários, sendo uma lista de comentários negativos e a outra de comentários positivos. No campo de comentário em análise estão todas as informações relevantes ao usuário que inseriu a frase, tais como: número de *likes*, seguidores, *retweets*, número de pessoas que o usuário segue, nome e o nível de criticidade (Figura 3).

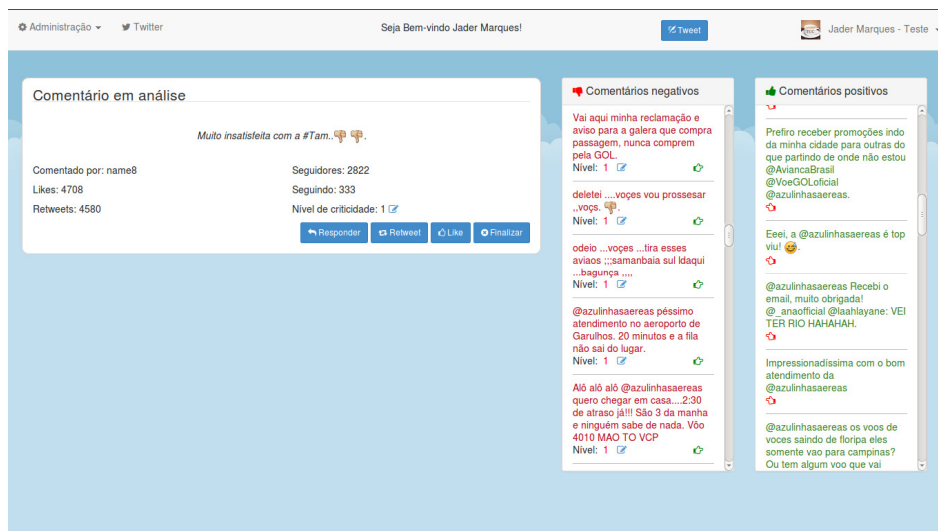


Figura 3. Tela de monitoramento.

Essas informações inerentes ao comentário permite ao analista ter uma visão do quanto pode ser impactante determinado comentário.

É possível ao analista curtir, responder, *retweetar* e finalizar o comentário. Caso o analista julgue que a análise realizada sobre o comentário não é procedente, o mesmo pode alterar de lista, trocando o comentário da lista de negativo para positivo, ou vice versa.

É concedida também a flexibilidade para o analista alterar o nível de criticidade conforme ele julgue adequado. O analista pode ainda inserir um comentário no Twitter.

5. Experimentos e Resultados

Inicialmente empresas da região foram procuradas para que pudessem utilizar a ferramenta para teste, e observou-se que as possíveis empresas não teriam um fluxo de dados suficiente para validar o comportamento do sistema. Com isso, buscou-se o contato com grandes empresas nacionais para oferecer a ferramenta para testes. A maioria destas empresas não permitiu que a aplicação tivesse acesso ao seu perfil do Twitter, em contrapartida algumas empresas permitiram que os comentários presentes em seu(s) perfil(s) fossem capturados de forma manual.

Diante disso, para teste da ferramenta capturou-se comentários reais de quatro empresas do ramo da aviação civil. O total de comentários coletados foi 516 (quinhentos e dezesseis), dos quais 292 continham frases negativas (56,6% do total), 155 continham frases positivas (30,8% do total) e 65 continham frases sem nenhum grau de sentimento, ou seja, frases neutras (12,6% do total) (Figura 4).

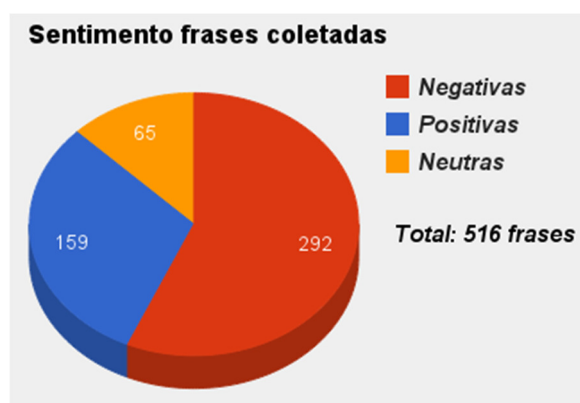


Figura 4. Frases coletadas.

Os resultados referentes à análise de frases de cunho negativo alcançaram uma acurácia de 84,6%, ou seja, de um total de 292 frases negativas 258 foram analisadas corretamente (Figura 5). Já a análise das frases de natureza positiva a ferramenta obteve uma precisão de acertos de 80,5%, sendo 128 de um total 159 frases analisadas (Figura 6).

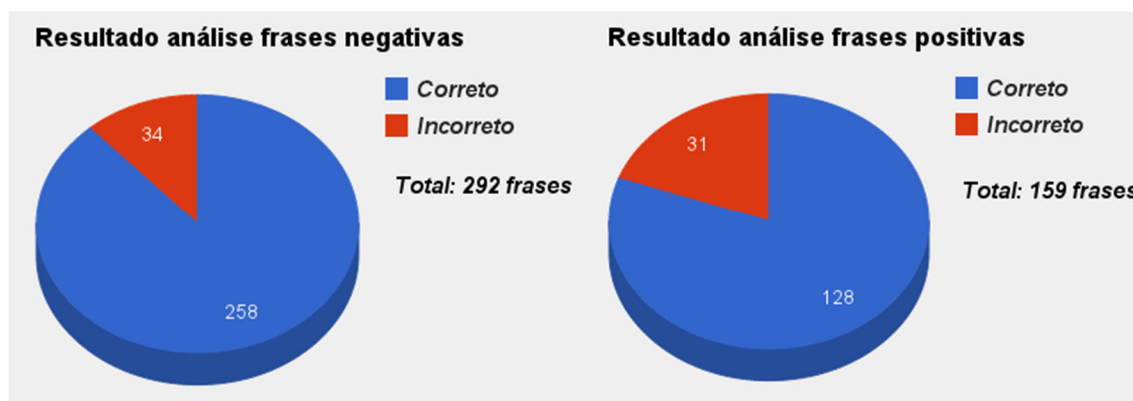


Figura 5. Frases negativas.

Figura 6. Frases positivas.

Dentro das 516 frases coletadas e analisadas, 65 eram neutras, ou seja, não expressavam sentimento algum. Dentro dessa análise 81,5% das frases foram analisadas corretamente (Figura 7).

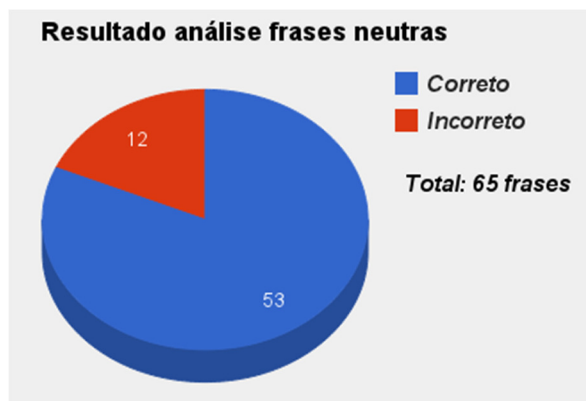


Figura 7. Frases neutras.

6. Conclusões

Este trabalho apresentou uma ferramenta que monitora o Twitter capturando os comentários e classifique-os em sua natureza positiva ou negativa, neste sentido algumas considerações devem ser feitas.

Primeiramente pensou-se em utilizar o Facebook, porém no decorrer do desenvolvimento observou-se que os comentários inseridos nessa mídia são muito vastos e com pouco objetivo, tornando assim a análise menos precisa. A escolha do Twitter foi acertada, pois ele possui uma API de fácil integração e seus comentários são em grande parte objetivos expressando claramente o que o usuário sente.

O fato de empresas locais não terem um fluxo considerável de informações no Twitter e também as grandes empresas contatadas não permitirem o acesso direto da aplicação gerou um fator complicado para o teste da ferramenta, o qual foi solucionado com a captura dos dados manualmente para realização dos testes.

No que se refere a análise de sentimentos dois fatores foram agravantes no desenvolvimento, que são o número elevado de erros de escrita da língua portuguesa, e o excesso de ironia nos comentários, deixando que uma grande parte dos comentários fossem analisados corretamente.

Os resultados obtidos com a ferramenta foram satisfatórios, levando em consideração o vasto vocabulário da língua portuguesa, o grande número de palavras ambíguas e os erros de português, a ferramenta conseguiu atingir um resultado bem expressivo.

7. Referências

- Bollen, J. et al (2009) "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena". Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, Spain.
- Cesário Júnior, J. M. (2010) "Sobre o Conceito de Processamento de Linguagem Natural", Trabalho acadêmico, Faculdade de Tecnologia, UNICAMP, Limeira.

- Chowdhury, G. C. (2003) "Natural Language Processing", Annual Review of Information Science and Technology, Vol. 37, p. 51-89.
- Correia, J. S. S. (2011) "Um Processo para Aquisição de Relações Taxonômicas de uma Ontologia". Dissertação, Universidade Federal do Maranhão, São Luís.
- Garcia, M. (2010) "Análise morfossintáctica para Português Europeu e Galego: Problemas, soluções e avaliação". *Linguamática*, 2(2): 59-67.
- Gomes, H. (2012) "Text Mining: Análise de Sentimentos na classificação de notícias". Tese de Mestrado, Universidade Nova Lisboa.
- Grigoletti, E. (2011) "Redes Sociais: O que mudou nas empresas?", Administradores.com, <http://www.administradores.com.br/noticias/tecnologia/redes-sociais-o-que-mudou-na-atuacao-online-das-empresas/46381/>, Julho.
- Koblitz, L. (2010) "Ambiente de Análise de Sentimentos Baseado em Domínio". Tese, Doutorado em Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro.
- Kopschitz, I. (2011) "Analista de redes sociais, um profissional do presente", <http://oglobo.globo.com/emprego/analista-de-redes-sociais-um-profissional-do-presente-2747776>, Setembro.
- Li, Y.-M. e Li, T.-Y. (2011) "Deriving Marketing Intelligence over Microblogs". Proceedings of 44th Hawaii International Conference On System Sciences (HICSS), pp. 1 –10.
- Liu, B. (2010) "Sentiment Analysis and Subjectivity", In Nitin Indurkha & F. J. Damerau (Eds.). Handbook of Natural Language Processing Second Edition. Boca Raton, FL: CRC Press, Taylor and Francis Group, Chicago, cap. 1, p.1-5.
- Liu, B. (2012) "Sentiment Analysis and Opinion Mining". Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Oliveira, N. (2013) "A importância das redes sociais para as empresas", <http://www.natanaeloliveira.com.br/a-importancia-das-redes-sociais-para-as-empresas/>, Maio.
- Pang, B. e Lee, L. (2008) "Opinion Mining and Sentiment Analysis". Foundation and Trends in Information Retrieval, Vol. 2 (1-2), pp. 1-135.
- Pang, B., Lee, L., e Vaithyanathan, S. (2002) "Thumbs up?: sentiment classification using machine learning techniques.", Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 79–86.
- Rodrigues, C. A. S. et al. (2010) "Mineração de Opinião e Análise de Sentimento", Trabalho acadêmico, Universidade Federal de Santa Catarina, Florianópolis. www.inf.ufsc.br/~alvares/INE5644/MineracaoOpinio.pdf.
- Sakaki, T. et al (2010) "Earthquake shakes Twitter users: real-time event detection by social sensors". Proceedings of the 19th International Conference on World Wide Web, (New York, NY, USA: ACM), pp. 851–860.