Avaliação de Algoritmos de Análise de Sentimentos em Tweets no Domínio da Copa do Mundo FIFA 2018

Patricia Feliciano[1], **Paulo Roberto Farah**[1]

[1]Departamento de Engenharia de Software
Universidade do Estado de Santa Catarina (UDESC)
Rua Dr. Getúlio Vargas, 2822 – 89.140-000 – Ibirama – SC – Brasil

patthyfeliciano@gmail.com, paulo.farah@udesc.br

Abstract. *Social networks have generated a large amount of data on the Web, so various methods and techniques have been proposed in the form of applications for monitoring and analysis of repercussion of brands, products relevant events such as the FIFA World Cup. This paper analyzes appropriate classification methods for the FIFA World Cup 2018 event domain on Twitter. It was collected more than 1 million of tweets, 4000 were labeled manually by a human and 400 thousand were classified. The results show that Opinion Finder algorithm had an almost perfect fit with kappa coeficient of 0,95 for the domain of the event.*

Resumo. *Redes sociais têm gerado um grande volume de dados na Web e vários métodos e técnicas vêm sendo propostos em forma de aplicações para monitoramento e análise de repercussão de marcas, produtos e eventos relevantes como a Copa do Mundo FIFA. Este artigo analisa métodos de classificação adequados para o domínio da Copa do Mundo FIFA 2018 no Twitter. Foram coletados mais de 1 milhão de tweets, 4000 foram rotulados manualmente por um humano e 400 foram classificados. Os resultados mostram que o algoritmo Opinion Finder obteve predição quase perfeita baseado no coeficiente kappa de 0,95 para o domínio do evento estudado.*

1. Introdução

A popularização e o aumento do uso das redes sociais na Internet transformaram o modo de interação entre pessoas e organizações. Por meio delas, diversos indivíduos possuem a liberdade de expressar suas opiniões sobre marcas, eventos, produtos, acontecimentos dentre vários assuntos que desejam publicar.

A expressividade dos usuários pode ser compartilhada com aspectos positivos, neutros ou até mesmo negativos como críticas, por exemplo. Logo, a rapidez de propagação e abrangência pode repercutir em impactos diretamente ligados a imagem de marcas e organizações empresariais, as quais geralmente não conseguem acompanhar a demanda de análise que todo conteúdo gerado requer ao mesmo tempo em que são postados.

Como nem sempre as repercussões das opiniões são positivas, estas por sua vez, podem comprometer e influenciar opiniões de demais indivíduos de forma negativa. Assim sendo, a análise destas informações pode servir vantajosamente como forma de extração de opiniões do público alvo acerca de seu produto ou serviço ofertado [Gomes 2013].

Nesse contexto, a mineração de textos disponibiliza um conjunto de técnicas capazes de automatizar o processo de extração e análise de sentimentos, de modo que as

organizações aproveitem os resultados obtidos para elaborar suas táticas de marketing em produtos e eventos, além de aperfeiçoar estratégias dentre outras possíveis aplicações [Tan 1999].

A Copa do Mundo FIFA é um evento esportivo muito popular. O campeonato foi disputado entre os dias 14 de junho e 15 de julho de 2018, por 32 seleções de diversos países [FIFA]. Atualmente é comum ver o crescimento exponencial da repercussão em torno de eventos de grande porte como esse.

A partir deste cenário, [Rezende 2005] afirma que: "Devido à incapacidade do ser humano de interpretar tamanha quantidade de dados, muita informação e conhecimento, possivelmente úteis, podem estar sendo desperdiçados, ficando ocultos dentro das Bases de Dados espalhadas pelo mundo". Assim sendo, a pergunta chave para definição do problema é: como obter e classificar sentimentos e opiniões expressados pelos usuários do Twitter a respeito da Copa do Mundo de 2018 utilizando o método mais adequado para este domínio?

O objetivo geral deste estudo é avaliar algoritmos na predição de sentimentos em tweets sobre a Copa do Mundo de 2018, de modo a descobrir o(s) método(s) de classificação com maior índice de concordância em predição sobre este evento. Para isto, este trabalho tem como objetivos específicos: (1) coletar dados textuais relacionados ao evento Copa do Mundo FIFA 2018; (2) rotular humanamente uma amostragem de tweets coletados; (3) categorizar sentimentos e opiniões extraídas da base de dados coletada; (4) validar o nível de acurácia dos métodos classificadores com a amostragem rotulada pelo ser humano.

Vistos os objetivos específicos, justifica-se o tema escolhido pela repercussão que eventos de grande abrangência vem a apresentar, como por exemplo a geração de opiniões e sentimentos, muitas vezes polêmicos em torno de assuntos relacionados. No esporte não é diferente, a popularidade do futebol no Brasil, país onde foi sediado o último Campeonato Mundial de Futebol FIFA 2014, aliado ao crescimento exponencial de adeptos e simpatizantes, foram fatores influentes na escolha da abordagem deste evento atual e popular.

O presente trabalho foi estruturado em quatro seções, sendo esta a primeira. A seção 2 contém alguns trabalhos relacionados a esse estudo. A seção 3 descreve a arquitetura e as etapas inerentes ao trabalho. A seção 4 apresenta os resultados obtidos. Por fim, a seção 5 possui a conclusão e os trabalhos futuros.

2. Trabalhos Relacionados

Esta seção apresenta trabalhos relacionados à mineração textual e análise de sentimentos em mídias sociais.

2.1. iFeel

Através de uma pesquisa em torno de técnicas e ferramentas existentes, um estudo relacionado apresenta um *benchmarking* entre 18 métodos de classificação de sentimentos [Araújo et al. 2016]. O projeto iFeel identifica e classifica sentimentos e opiniões em textos estruturados e não estruturados. Os autores explicam que um método que apresentou desempenho eficaz em um contexto de 'política', não era tão eficaz quando aplicado a um

contexto de 'esportes', exemplo meramente ilustrativo. Isso se dá em virtude de contextos de treinamento realizados, dicionários léxicos distintos, presença de expressões populares em torno de um assunto específico, ou seja, vários fatores podem refletir no desempenho de um método classificador.

Os autores comparam os seguintes algoritmos: Opinion Lexicon, Sentistrength, Socal, Happiness Index, Sann, EmoticonsDS, Sentiment, Stanford, Afinn, Mpqa, Nr-chashtag, Emolex, Emoticons, Sasa, Panast, Vader e Umigon. De acordo com as métricas aplicadas pelos autores, os cinco métodos que mais se destacaram foram: Sentistrength, Afinn, Opinion Lexicon, Umigon e Vader [Benevenuto et al. 2015].

2.2. Copa do Mundo FIFA 2014

Este trabalho analisa e compara o comportamento das torcidas fora dos campos. Além disso, toma como consideração final a concepção do impacto resultante da análise textual para analisar e auxiliar no entendimento da dinâmica desse comportamento [Kim et al. 2015].

Nessa pesquisa 790.744 usuários do Twitter tiveram seu comportamento acompanhado e suas mensagens registradas e analisadas tanto antes quanto durante o evento. Nesse sentido, foram analisados padrões temporais de volumes de mensagens, tópicos repercutidos, compartilhamento de postagens (retweets), dentre outros.

Os autores pesquisaram três questões principais, sendo que a primeira analisou se o comportamento dos participantes apresentou variação nos momentos de transmissão comparado à momentos normais em que não estava sendo transmitido o evento. A segunda questão tratou de como a diversidade de tópicos muda durante o evento. Por fim, o terceiro tópico analisado relacionou como os tweets variaram de acordo com a proximidade dos países com o evento FIFA 2014, bem como o papel dos usuários multilíngues sobre as diferenças geográficas e culturais dentro das comunidades online.

Neste contexto, relata-se a abordagem de busca utilizada, onde traduziu-se o termo "worldcup2014" para vários idiomas para filtrar os tweets coletados, bem como a utilização de abreviaturas de nomes de seleções que estariam disputando os jogos, como por exemplo: (#BRAvsGER).

3. SAMS

Esta seção apresenta o projeto Sistema de Análise e Monitoramento de Sentimentos (SAMS). Inicialmente é descrita a arquitetura dos processos incorporados ao presente trabalho, seguido das etapas de desenvolvimento.

3.1. Arquitetura Geral

A arquitetura geral do SAMS é apresentada na Figura 1 a qual contém os módulos e suas respectivas ferramentas adotados na pesquisa.

Inicialmente, a arquitetura possui um módulo coletor (1) que tem como objetivo coletar tweets em tempo real com base nas palavras-chaves inicialmente especificadas. É responsável pela persistência na base de dados dos tweets coletados (2) para, em seguida, serem exportados com auxílio da ferramenta de exportação (3). Logo acontece a retroalimentação, onde identifica-se a frequência das palavras que mais aparecem na coleta e

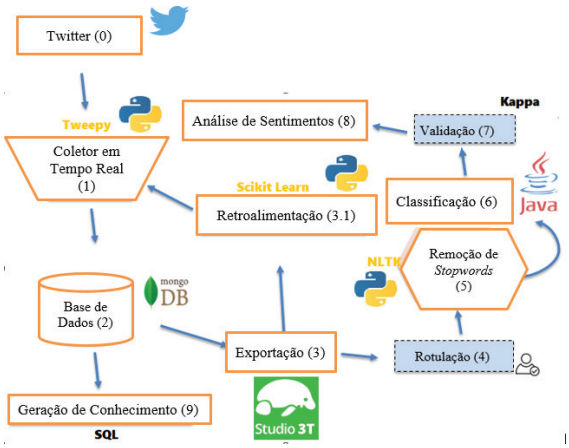


Figura 1. Arquitetura geral do SAMS.

assim, estas serem analisadas e inseridas no filtro para robustez do coletor. Deste modo, a ferramenta Robo3T, a qual basicamente permitiu a elaboração de scripts para geração de conhecimento, bem como a exportação dos dados em formato CSV contribui para dois fluxos desta arquitetura.

A partir das planilhas CSV exportadas foi possível efetuar os tratamentos necessários; como a remoção de *stopwords* (5) para em seguida submeter estes dados aos métodos de classificação (6). Feito isto, acontece a validação dos resultados rotulados versus os resultados provenientes da classificação automatizada (7). Nesta fase de validação foram geradas as matrizes de confusão e o coeficiente Kappa, ambos úteis na identificação dos métodos com a predição mais assertiva para a Copa do Mundo FIFA 2018. Na sequência, acontece a análise de sentimentos (8). Ao final da análise de sentimentos foi realizado a geração de conhecimento (9), esta fase apesar de ser exposta como a última não possui relação direta da fase(8), pois trata-se de uma exploração da base de dados coletada a qual envolve análises via SQL representadas graficamente.

Neste contexto, [Landis and Koch 1977] classificam a força da concordância através de faixas analisadas sobre o índice calculado. A Tabela 1 apresenta a classificação do índice Kappa. Pode-se avaliar o nível de concordância que um coeficiente Kappa Cohens pode assumir, partindo da insignificância representada pelo valor zero passando por várias faixas até o valor 1, o qual caracteriza uma excelente concordância em um cenário quase perfeito.

Tabela 1. Classificação do Índice Kappa

Valor	Interpretação		Valor	Interpretação
< 0	Ausência de concordância		0,41 – 0,60	Moderada
0,00 – 0,20	Mínima		0,61 – 0,80	Substancial
0,21 – 0,40	Razoável		0,81 – 1,00	Quase perfeita

3.2. Etapas

Nesta seção serão apresentadas as etapas de desenvolvimento de estudo, deste a coleta até os resultados obtidos. Para a elaboração do presente trabalho foram realizadas as seguintes etapas: coleta, rotulação, categorização, validação e análise de resultados.

3.2.1. Coleta

O Twitter foi escolhido por ser uma plataforma com um grande número de usuários ativos. Além disto, a abrangência mundial, tanto desta rede social quanto do evento FIFA 2018 são fatores que podem instigar usuários a compartilhar ideias constantemente. O Twitter disponibiliza uma API que possibilita de maneira prática a exploração de informações textuais e atributos públicos contidos nos tweets postados.

Diante deste cenário, a metodologia utilizada foi a coleta de dados, a validação de classificadores, comparados ao treinamento humano, a qual preocupa-se em analisar a acurácia dos métodos estudados e sua eficácia na resolução dos problemas propostos diante de características do evento estudado. A extração de dados textuais iniciou-se em abril de 2018, sendo executada até início de junho em dias variados com intervalos no período de extração. Foram coletados atributos públicos como: tweet, coordenadas, localização, país, idioma, data de criação da conta do usuário, data da postagem, quantidade de amigos e quantidade de seguidores, cor de fundo da conta do usuário, tipo de dispositivo que originou a postagem coletada e persistidos em uma coleção no MongoDB.

3.2.2. Rotulação

Nesta etapa de desenvolvimento foram rotulados manualmente 2 mil tweets aleatórios em português e 2 mil também aleatórios em inglês. No momento de rotulação, foram consideradas as intenções que o usuário teve ao expressar-se no Twitter. De modo que a opinião particular do rotulador em polêmicas presentes no texto não foi considerada.

Os tweets com apenas links, datas, marcações de outros usuários, ambiguidade de interpretação, ironias, propagandas, ou que não expressavam nenhum sentimento plausível bem como opiniões sobre assuntos relacionados ao evento foram considerados neutros (0). Em contrapartida, tweets que claramente expressavam sentimentos bons, opiniões favoráveis, foram categorizados com positivo (1). De forma que os demais tweets foram vistos como negativos (-1), sendo eles portadores de características ruins, triste, adversos ou mesmo vocabulários obscenos, tanto no idioma inglês quanto no português.

3.2.3. Validação de Classificadores

Após a rotulação humana em uma significativa amostragem da base de dados, tornou-se viável a validação dos métodos, no propósito de identificar quais os melhores métodos dentre os 18 disponibilizados em [Benevenuto et al. 2015, Araújo et al. 2016]. Os resultados obtidos com esta avaliação, não são capazes de declarar uma verdade única ou prover um *benchmarking* completo, mas sim, indicar através de métricas estatísticas quais os melhores métodos para o evento específico idealizado no presente trabalho. Visto que, conforme [Benevenuto et al. 2015], ainda não se encontrou um método classificador com desempenho acima de todos os demais existentes. Por isto a importância de uma validação como esta antes de aplicar qualquer método em qualquer contexto e analisar sentimentos às cegas.

Antes de submeter os arquivos para o iFeel realizou-se um pré-processamento do

texto, o qual consistiu na remoção de caracteres especiais, exceto o "#" que caracteriza uma *hashtag* na maioria de suas aparições. Nesse sentido, também foram removidas palavras sem relevância, conhecidas como *stopwords*, sendo para o descarte destas palavras utilizou-se a plataforma NLTK (Natural Language Toolkitk), empregada no Python para trabalhar com a linguagem humana. Tudo isto para melhorar o desempenho dos algoritmos de classificação textual a serem utilizados na sequência.

Foram submetidos os mesmos 4 mil tweets rotulados na subseção anterior aos cinco melhores métodos do iFeel (Sentistrength, AFINN, OpinionLexion, Umigon e Vader) conforme elencados pelo *benchmarking* de [Araújo et al. 2016]. Visto isto, para geração das matrizes de confusão de todos os métodos analisados utilizou-se as tecnologias já existentes provenientes da biblioteca de aprendizado de máquina Scikit Learn que consequentemente deram acesso à funções nativas desta API: confusion_matrix e cohen_kappa_score.

4. Resultados

Esta seção apresenta os resultados obtidos com o presente trabalho. A começar pelos principais resultados provenientes das matrizes de confusão, as quais viabilizaram o cálculo do coeficiente Kappa (κ) e o índice de concordância desenvolvido para validação dos métodos de classificação que foram utilizados.

Inicia-se a apresentação dos resultados pela abordagem das seis principais matrizes de confusão geradas para a base de dados em inglês sobre os métodos de classificação utilizados neste estudo que mais se destacaram perante aos demais, vide Tabela 2.

Tabela 2. Matrizes de Confusão

	OpinionFinder				Umigon				SentiStrength		
	+	N	-		+	N	-		+	N	-
+	467	16	0		468	15	0		471	12	0
N	7	1478	3		58	1413	17		85	1337	66
-	0	8	21		0	22	7		0	8	21
	OpinionLexicon				Vader				Affin		
	+	N	-		+	N	-		+	N	-
+	471	12	0		468	15	0		471	10	2
N	268	1198	22		230	1241	17		522	911	5
-	0	16	13		1	24	4		3	18	8

Observa-se, a partir da Tabela 2, a quantidade de verdadeiros e falsos acertos que cada um dos seis métodos exibidos acima apresentou perante a rotulação humana sobre a amostragem em inglês. Deste modo, caracteriza-se a diagonal principal como os verdadeiros acertos, ou seja, o conjunto de dados foi rotulado positivo tanto pelo método o quanto pelo rotulador, o mesmo aplica-se às duas outras categorias negativo e neutro.

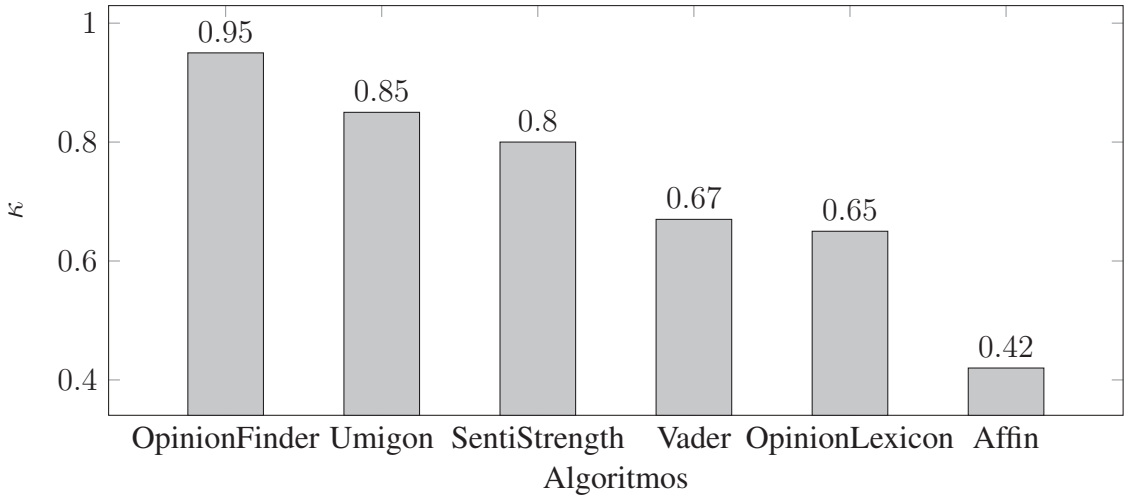
Além disso, nota-se que a quantidade de acertos verídica, diagonal principal é maior para o método *Opinion Finder* do que para os demais, principalmente o *Affin* que resultou em muitas divergências nas demais posições da matriz.

Na sequência, ao analisar o índice de concordância, e observou-se que, para a base

de dados em português, nenhum dos 18 métodos do iFeel apresentou classificação do índice Kappa (κ) igual ou superior a *0,40* o que indica que nenhum método de classificação resultou em predições no mínimo moderadas em relação a este contexto e idioma.

Portanto, optou-se por aplicar as demais etapas de desenvolvimento somente aos tweets coletados em inglês, visto que para este idioma submetido aos mesmos métodos obteve-se resultados muito mais satisfatórios, cujo Kappa dos seis melhores métodos variou entre *0,42* e *0,95*, onde 1 corresponde a predição excelente. Em seguida, a Figura 2 representa graficamente os principais resultados da validação através do índice de concordância Kappa.

Figura 2. Resultado da comparação do índice κ entre os algoritmos avaliados.



Pode-se observar que, dentre os seis métodos com maior índice de concordância estão presentes os mesmos cinco métodos que mais se destacaram no *benchmarking* de [Araújo et al. 2016]. Entretanto, o segundo ponto observado é que *Opinion Finder* foi o método que melhor se destacou com um coeficiente de *0,95* o que caracteriza uma predição quase perfeita.

5. Conclusão e Trabalhos Futuros

O presente trabalho teve como objetivo principal a classificação de algoritmos para análise de sentimentos no domínio da Copa do Mundo FIFA 2018 no Twitter. Logo, através de etapas de desenvolvimento e validações efetuadas conclui-se que o objetivo foi alcançado. Inicialmente foram expostos cinco objetivos específicos, os quais também foram contemplados. Este estudo de caso coletou dados textuais relacionados ao evento almejado. Além disto, rotulou-se uma amostragem significativa da base de dados coletada do Twitter. Na sequência foram categorizados os sentimentos e opiniões presentes na base de dados extraída para amostragem rotulada de modo a utilizar os classificadores do iFeel.

Foram validados os níveis de acurácia dos classificadores comparado a rotulação humana. E, por fim, foi classificada a base de dados coletada no idioma inglês, visto que durante a etapa de validações este foi o idioma com melhores resultados apresentados.

Aliado às funcionalidades, desenvolvimento e validações estatísticas efetuadas, SAMS é um estudo de caso capaz de identificar dentre vários métodos, qual o melhor

método para um contexto específico. Isto significa que as predições dos classificadores utilizados não acertaram por um simples acaso, mas sim, porque cientificamente apresentaram-se na faixa de classificação mais próxima a predição perfeita, caracterizando assim um diferencial de suma importância deste trabalho perante aos estudos correlatos explanados.

Em resumo, mais de 1 milhão de tweets foram coletados, juntamente com os 4 mil tweets rotulados humanamente, além de, 400 mil tweets classificados, bem como as matrizes de confusão e acurácia geradas para cada um dos vários métodos estudados. Por fim, o presente trabalho também possibilitou a geração do índice de concordância dos classificadores e validações efetuadas sobre o estudo e caracterizaram os principais resultados obtidos sobre os objetivos previamente estipulados.

Em trabalhos futuros pretende-se caracterizar as postagens coletadas a respeito da Copa do Mundo FIFA 2018. Além disso, comparar os métodos estudados em outros eventos relevantes de tamanha abrangência.

Referências

Araújo, M., Diniz, J. P., Bastos, L.and Soares, E., Júnior, M., Ferreira, M., Ribeiro, F., and Benevenuto, F. (2016). ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. In *Proceedings of the International AAAI Conference on Web-Blogs and Social Media*, Cologne, Germany.

Benevenuto, F., Ribeiro, F., and Araújo, M. (2015). Métodos para análise de sentimentos em mídias sociais. In *Brazilian Symposium on Multimedia and the Web (Webmedia)*, Manaus, Brasil.

FIFA. Site oficial do evento copa do mundo fifa 2018. `https://www.fifa.com/worldcup`. Acesso em junho de 2018.

Gomes, H. J. C. (2013). Análise de sentimentos na classificação de notícias. In *Proceedings of the 8th Iberian Conference on Information Systems and Technologies (CISTI)*, Lisboa, Portugal.

Kim, J. W., Kim, D., Keegan, B., Kim, J., Kim, S., and Oh, A. (2015). Social media dynamics of global co-presence during the 2014 fifa world cup. In *CHI'15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Republic of Korea. ACM New York, NY, USA ©2015.

Landis, J. and Koch, G. (1977). An application of hierarchical kappa type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33(2):363–374.

Rezende, S. O. (2005). Mineração de dados. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, São Leopoldo, Brasil.

Tan, A. H. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD Workshop on Knowledge Disocovery from Advanced Databases*, Beijing, China.