CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
Engenharia de Computação

Andressa Oliveira Souza

IDENTIFICAÇÃO E CLASSIFICAÇÃO DE USUÁRIOS INFLUENTES NO TWITTER POR ÁREAS DE INFLUÊNCIA

Timóteo
2015

Andressa Oliveira Souza

IDENTIFICAÇÃO E CLASSIFICAÇÃO DE USUÁRIOS INFLUENTES NO TWITTER POR ÁREAS DE INFLUÊNCIA

Monografia apresentada ao Curso de Engenharia de Computação do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Engenheira de Computação.

Orientador: Luciano Nascimento Moreira

Timóteo
2015

AGRADECIMENTOS

Agradeço a Deus e a Nossa Senhora, por terem me dado a oportunidade de adquirir conhecimento e condições para aproveitá-la.

Agradeço aos meus pais, Marcos e Silvia, e a minha irmã, Amanda, pelo amor, incentivo e por poder contar sempre com eles.

Agradeço ao meu esposo e amigo Leonardo, pelo companheirismo e pelo apoio.

Agradeço ao meu filho Nicolas, minha maior alegria e motivação.

Agradeço o professor Luciano Nascimento Moreira, pela dedicação e auxílio no desenvolvimento deste trabalho.

E por fim, agradeço aos meus colegas de curso e todos que de alguma forma me ajudaram no desenvolvimento deste trabalho.

RESUMO

As redes sociais estão cada vez mais presentes no cotidiano das pessoas. Através delas as pessoas buscam e emitem opiniões, sejam elas sobre algum acontecimento do dia, posicionamento político, ou até mesmo sobre algum produto. Com isso, algumas empresas e políticos vêm contratando para fins de marketing pessoas que não são necessariamente famosas, mas são influentes em um grupo de clientes ou eleitores potenciais. Assim, este trabalho identificou usuários influentes de um determinado grupo, por meio do número de menções, retweets e seguidores e categorizou as áreas de influência destes usuários, baseando-se na classificação do texto de suas postagens. Em cima dos rankings gerados, foi realizada a classificação das postagens por meio do algoritmo de Naive-Bayes. As postagens foram classificadas como pertencentes às categorias: esporte, política ou outros assuntos. Baseando-se nos resultados obtidos nos treinamentos e classificações com maior e menor percentual de postagens classificadas corretamente, foi calculado o percentual de pertencimento dos usuários a cada uma das categorias. A categoria política apresentou o maior coeficiente de determinação médio que determina a correlação entre o percentual de pertencimento dos usuários obtido por esta categoria e o valor real, com aproximadamente 0.95871 e o desvio padrão de 0.02322. A categoria esporte obteve o coeficiente de determinação médio de 0.46323, com desvio padrão de 0.19478. A categoria outros assuntos obteve o coeficiente de determinação médio de 0.47085 e o desvio padrão de 0.19249. Conclui-se então que o processo de classificação das postagens realizado neste trabalho foi eficaz apenas para as postagens da categoria política.

Palavras-chave: Classificação de texto. Naive-Bayes. Redes sociais. Usuários influentes. *Twitter*.

SUMÁRIO

1 INTRODUÇÃO

Segundo Rocha e Alves (2010), nos últimos anos as redes sociais vêm conquistando espaço na vida das pessoas, se adequando aos diversos assuntos e gostos. Um estudo feito em 27 países desenvolvido pelo Ibope (2010), concluiu que o Brasil estava em 10ª colocação no ranking dos países que mais utilizavam as redes sociais. De acordo com Conecta (2014), em um estudo realizado com internautas entre 15 e 32 anos, concluiu-se que 90% dos entrevistados acessam as redes sociais. Neste mesmo estudo, tem-se que as cinco redes sociais mais populares são: *Facebook* (onde 96% possuem perfil), *YouTube* (79%), *Skype* (69%), *Google+* (67%) e *Twitter* (64%).

Conecta (2014) também constatou que 41% dos entrevistados concordam parcialmente ou totalmente que “para encontrar informações confiáveis sobre marcas, produtos e serviços tem que buscar outras fontes de informação além das próprias marcas e empresas” contra 20% que discordam parcialmente ou totalmente e “53% afirmam que falam das marcas e produtos com amigos e familiares a fim de trocar ideias e experiências... ‘e, naturalmente, influenciar... ’”. Com isso, pode-se concluir que as pessoas buscam e emitem informações sobre marcas, produtos e serviços e possuem em alguma proporção a capacidade de influenciar alguém.

> Identificar quem são os usuários influentes, e como a informação relevante se propaga nas redes sociais são tarefas que, se bem resolvidas, trazem conhecimento estratégico para empresas de marketing, campanhas políticas e estudos sociológicos. (VALIATI *et al*., 2012, p.2)

A identificação dos usuários influentes e a categorização da(s) área(s) que esta pessoa exerce influência (Ex: música, política, moda, etc.) possibilita parcerias de marketing menos usuais e mais eficientes, o que pode se tornar um diferencial para empresas.

Assim, este trabalho busca responder o seguinte problema: seria possível categorizar áreas de influência de usuários influentes por meio da análise das suas postagens?

1.1 Justificativa

Diante das várias possibilidades de uso como aquelas apresentadas na seção anterior, vários trabalhos vêm sendo desenvolvidos para a identificação de usuários influentes nas diversas redes sociais existentes. Porém, os trabalhos encontrados para identificação de usuários influentes não categorizam as áreas de influência, considerando que estes usuários são influentes em diversas áreas.

1.2 Objetivos

Este trabalho possui como objetivo geral a criação de um método para identificar usuários da rede social *Twitter* que exercem algum tipo de influência aos seus seguidores e que se baseando no conteúdo de suas postagens, classifique de forma automática em quais categorias esses usuários são influentes.

Para atingir o objetivo geral, este trabalho também possui os seguintes objetivos específicos:

1. produzir uma coleta de dados de rede social;
2. aplicar um método para identificação e ranqueamento dos usuários influentes nos dados coletados;
3. aplicar um algoritmo classificador nos dados melhor posicionados no *ranking*;
4. comparar os resultados da classificação automática com uma classificação feita manualmente.

1.3 Estrutura Textual

Este trabalho encontra-se organizado da seguinte forma:

- No capítulo 2 são apresentados alguns trabalhos recentes desenvolvidos para identificar usuários influentes em redes sociais.
- No capítulo 3 são apresentados os trabalhos desenvolvidos utilizando classificadores de texto em postagens de redes sociais.
- O capítulo 4 apresenta a fundamentação teórica necessária para compreensão deste trabalho.

- O capítulo 5 apresenta os procedimentos metodológicos utilizados na realização das etapas que compõem este trabalho.
- No capítulo 6 são apresentados os resultados obtidos.
- No capítulo 7 apresenta as conclusões deste trabalho, assim como identifica alguns possíveis trabalhos futuros.

2 USUÁRIOS INFLUENTES EM REDES SOCIAIS

Com o grande número de redes sociais disponíveis atualmente, têm-se também vários trabalhos desenvolvidos para identificar seus usuários mais influentes. Uma das redes sociais mais exploradas nesses trabalhos é o *Twitter*.

Valiati *et al*. (2012) buscam identificar usuários influentes e conteúdo relevante se utilizando de uma técnica que o autor define como uma extensão da técnica de *PageRank* (PAGE *et al*., 1998).

No contexto de redes sociais, adaptações do algoritmo de *PageRank* vêm sendo utilizadas para identificar usuários influentes. O *TwitterRank* (WENG *et al.*, 2010) surgiu para medir a influência de usuários no *Twitter*. Segundo Weng *et al.* (2010), um grafo orientado *D(V, A)* é formado entre o relacionamento de usuários e "*seguidores*". *V* é o conjunto de vértices formado por todos os usuários. Existe uma aresta entre dois usuários quando há uma relação de seguir entre eles, a aresta é direcionada do seguidor para o seguido. Um modelo de navegação sobre o grafo *D* executa o *TwitterRank* da seguinte maneira: um navegador aleatório visita cada usuário com uma certa probabilidade, seguindo a aresta em *D*. Weng *et al.* (2010) ainda ressalta que o *TwitterRank* se diferencia do *PageRank* pelo fato do navegador aleatório executar uma caminhada aleatória por tópicos específicos, ou seja, a probabilidade de transição de um usuário para outro é específica por tópicos.

Em seu trabalho, Bigonha (2012) identificou usuários influentes baseando-se em três elementos: relacionamento com os vizinhos, polaridade das suas opiniões e as características textuais dos seus *tweets*, por meio do método *Sentiment-based Influence Detection on Twitter (SaID)*, baseado no *PageRank*. Esse método foi proposto para identificar usuários influentes no *Twitter* baseando-se em sentimento. Segundo Bigonha (2012), o método é dividido em três fases: pré-processamento, extração de características e pontuação de influência. Bigonha (2012) diz que a fase de pré-processamento consiste em definições de consultas e tópicos percorrendo as postagens, filtrando o conteúdo, identificando os usuários e extraindo os dados. A fase de extração de características é definida por Bigonha (2012) como o processamento das métricas, onde o *SaID* processa o conteúdo, as interações e conexões entre os usuários e analisa o sentimento das postagens. Bigonha (2012)

ainda diz que a fase de pontuação de influência combina essas métricas e gera uma pontuação.

Gardelha (2013), por sua vez, busca em seu trabalho identificar influenciadores através de traços de personalidade, utilizando algoritmos de regressão. Gardelha (2013) diz ainda que é possível caracterizar influenciadores como indivíduos emocionalmente estáveis, extrovertidos, organizados e criativos.

Em seu trabalho feito para a rede social *Twitter*, Cha *et al*. (2010) apresentam uma comparação de três medidas de influência:

1. número de seguidores, onde se têm os usuários que possuem maior público;
2. número de *retweets*, onde se têm os usuários que possuem maior capacidade de ter sua postagem difundida;
3. número de menções, onde se têm os usuários que conseguem obter uma maior interação com outros usuários;

Cha *et al*. (2010) ainda conclui que popularidade (um maior número de seguidores) pode não necessariamente implicar em influência.

3 CLASSIFICAÇÃO DE TEXTO EM POSTAGENS

Os algoritmos de classificação vêm sendo utilizados para mineração de textos em redes sociais. Por meio deles, é possível analisar comportamentos e opiniões de usuários.

Em seu trabalho, Santos *et al.* (2010) analisam o quanto os usuários da rede social *Twitter* expressam suas opiniões sobre um certo produto. Santos *et al.* (2010) utilizaram o algoritmo classificador *SVM* para classificar as postagens como neutras ou opinativas.

As técnicas de classificação de texto são utilizadas também para a análise de sentimento em postagens. Liu (2012, p.7) define análise de sentimento como o campo que estuda e analisa a opinião das pessoas, sentimentos, avaliações, atitudes e emoções com relação a entidades como produtos, serviços, organizações, entre outros.

Ferreira (2010) utiliza em seu trabalho técnicas de classificação de texto para análise de sentimento em postagens do *Twitter* classificando as opiniões em positivas, negativas, neutras, ou não classificáveis. Ferreira (2010) adaptou uma técnica de classificação proposta para fóruns de discussão para utilizá-la em postagens. Ferreira (2010) utilizou para a classificação das postagens o algoritmo SVM, fazendo validação cruzada de 10 vezes e como função do núcleo do SVM utilizou a função de base radial com o algoritmo de otimização mínima sequencial.

Nascimento *et al.* (2012) realizou a análise de sentimentos de postagens com foco em notícias, utilizando-se dos classificadores: *Unigrama*, *Octograma* e *Naive-Bayes*. Nascimento *et al.* (2012) chegou a conclusão de que a classificação com o algoritmo *Naive-Bayes* obteve melhor resultado ao categorizar textos com características de *microblog* e fez a observação de que o classificador *Octograma* obteve melhor resultado do que o *Unigrama*.

Bigonha (2012), também utiliza em seu trabalho a análise de sentimentos para verificar a polaridade das opiniões emitidas pelos usuários sobre um determinado tópico.

Neste trabalho não utilizaremos a Análise de Sentimentos nas postagens, mas utilizaremos técnicas de classificação de texto semelhantes para identificar as áreas relacionadas às postagens dos usuários e páginas influentes.

4 APRENDIZADO DE MÁQUINA

Pardo e Nunes (2002) dizem que as técnicas de Aprendizado de Máquina (AM) vêm sendo utilizadas em vários ramos da computação, tais como: reconhecimento de imagens, sistemas baseados em conhecimento, roteamento de redes e processamento de textos, com resultados satisfatórios.

Mitchell (1997) diz que um programa de computador aprende com a experiência *E* no que diz respeito a alguma classe de tarefas *T* e medida de desempenho *D*, se o desempenho em tarefas em T, tal como medido por *D*, melhora com experiência *E*.

Murphy (2012) define o aprendizado de máquina como um conjunto de métodos que podem detectar padrões em dados e posteriormente, utiliza estes padrões para prever dados futuros. “As técnicas de AM empregam um princípio de inferência denominado indução, no qual obtêm-se conclusões genéricas a partir de um conjunto particular de exemplos” (SCHIMITT, 2013, p.21).

4.1 Conceitos de aprendizado de máquina

Santos (2012) diz que apesar de haver uma grande variedade de métodos de aprendizagem de máquina diferentes, alguns termos são comuns em sua grande maioria, são eles:

- Atributo: um atributo representa uma característica ou aspecto de um exemplo. Podem ser discretos ou contínuos. Atributos discretos contêm um número finito de valores, enquanto os contínuos podem assumir um número infinito de valores (MONARD, 2003; FACELLI *et al*, 2011, citado por Santos(2012)).

- Classe ou rótulo: descreve o conceito-meta, ou seja, o conceito que se deseja aprender para tornar viável a realização de predições a seu respeito.

- Instância ou exemplo: uma instância ou exemplo descreve um objeto através de um vetor com valores de seus atributos (MONARD, 2003, citado por Santos (2012)).

- Conjunto de dados, instâncias ou exemplos: um conjunto de dados é formado por um número de instâncias e seus respectivos atributos.

- Conjunto de treinamento: utilizado como entrada pelos algoritmos de aprendizado. Serve de base para construção de modelos ou hipóteses.

- Conjunto de teste: utilizado para avaliar o modelo construído.

- Conjunto de validação: utilizado para realizar ajustes no modelo construído pelo algoritmo de aprendizado.

- Indutor: algoritmo de aprendizado que utiliza processo indutivo para a criação do modelo.

- Modelo, classificador ou hipótese: dado um conjunto de dados de treinamento, um indutor ou algoritmo de aprendizado gera como saída um classificador (ou hipótese) de forma que, dado uma nova instância, ele possa predizer sua classe (MONARD, 2003, citado por Santos (2012)).

- *Overfitting* (superajustamento): o *overfitting* ocorre quando o classificador gerado é muito específico para o conjunto de exemplos de treinamento e a maior parte das instâncias que não pertence ao conjunto de dados utilizado durante a aprendizagem é classificada erroneamente.

- *Underfitting*: o *underfitting* ocorre quando o classificador gerado ajustou-se muito pouco ao treinamento, incapacitando-o de classificar corretamente tantos as instâncias do conjunto de treinamento, quanto às do conjunto de teste (MONARD, 2003, citado por Santos (2012)).

- Acurácia: a acurácia de um classificador é uma medida de desempenho obtida para uma determinada tarefa. Ela pode ser calculada de acordo com a taxa de predições corretas (precisão) ou incorretas (taxa de erro).

4.2 Tipos de aprendizado de máquina

Schimitt (2013) aponta que as técnicas de aprendizado de máquina são divididas em dois tipos principais: não supervisionado e supervisionado.

4.2.1 Aprendizado Não Supervisionado

Pardo e Nunes (2002) dizem que no aprendizado não supervisionado, o algoritmo de aprendizado é encarregado de tentar agrupar os dados de acordo com suas características da melhor maneira possível. Costa (1999) ressalta que diferentemente dos sistemas supervisionados, que são treinados com conjuntos rotulados de padrões, onde se têm no rótulo a categoria pertencente do padrão, a única informação utilizada nos sistemas não supervisionados são os próprios padrões.

Santos (2012) aponta como métodos de aprendizado não supervisionado as Redes Neurais do tipo Mapa Auto-Organizáveis (SOM, do inglês, *Self- Organizing Maps*), o k-Médias (do inglês, *k-Means*) e o Agrupamento Hierárquico.

Como este trabalho buscou descobrir o pertencimento de postagens de rede social às categorias pré-definidas, não foi utilizado o tipo de aprendizado não supervisionado.

4.2.2 Aprendizado Supervisionado

Pardo e Nunes (2002) definem o aprendizado supervisionado quando cada instância está associada a sua classificação, que o algoritmo de aprendizado deve aprender a realizar.

Santos (2012) diz que, durante um treinamento, para apresentar um exemplo ao algoritmo de aprendizado, ele deve possuir os atributos que representam as

características do domínio e um atributo que representa a classe a qual este exemplo pertence.

No aprendizado supervisionado “[...] o objetivo é aprender um mapeamento de x entradas para y saídas [...]” (MURPHY, 2012, p.3, tradução nossa).

Alguns métodos de aprendizado supervisionado são: o Aprendizado baseado em instâncias, o Aprendizado Bayesiano, as Árvores de decisão e regras; as Máquinas de Vetores Suporte e as Redes neurais artificiais.

Neste trabalho, optou-se por utilizar um método de aprendizado supervisionado, o aprendizado bayesiano. Este método foi escolhido por ser simples e proporcionar um baixo custo computacional.

4.2.2.1 Aprendizado Bayesiano

Mitchell (1997) define que os algoritmos de aprendizagem bayesianos são aqueles que calculam probabilidades explícitas para as hipóteses, e diz que eles fornecem abordagens práticas para certos tipos de problemas de aprendizagem.

Pardo e Nunes (2002) citam como vantagens deste tipo de aprendizagem a possibilidade de se incluir nas probabilidades calculadas o conhecimento de domínio que possa ter, e a capacidade das classificações serem baseadas em evidências fornecidas, podendo aumentar ou diminuir as probabilidades das classes a serem observadas em uma nova instância que se quer classificar.

4.2.2.1.1 Classificador de Naive Bayes

Pardo e Nunes (2002) descrevem o classificador Naive-Bayes como baseado no Teorema de Bayes para o cálculo das probabilidades necessárias para a classificação. Sendo assim, dada uma nova instância $A=a_1, a_2, \ldots a_n$, deseja-se predizer sua classe:

$$P(classe|A) = \frac{P(classe|A) \times P(classe)}{P(A)} \quad (1)$$

Pardo e Nunes (2002) também dizem que como $A= a_1, a_2, \ldots a_n$, tem-se:

$$P(classe|a_{1.} \dots a_n) = \frac{P(a_{1.} \dots a_n|classe) \times P(classe)}{P(a_{1.} \dots a_n)} \quad (2)$$

“Para calcular a classe mais provável da nova instância, calcula-se a probabilidade de todas as possíveis classes e, no fim, escolhe-se a classe com a maior probabilidade como rótulo da nova instância.” (PARDO; NUNES, 2002, p.5)

Pardo e Nunes (2002) dizem ainda que em termos estatísticos isso se equivale a maximizar $P(classe|a_{1.} \dots a_n)$ e para isso, deve-se maximizar o numerador $P(a_{1.} \dots a_n|classe) \times P(classe)$ e minimizar o denominador $P(a_{1.} \dots a_n)$. Pardo e Nunes (2002) mostra que como o denominador $P(a_{1.} \dots a_n)$ é uma constante, por não depender da variável classe que se está procurando, pode-se anulá-lo, resultando na fórmula abaixo:

$$\text{argMax}\, P(classe|a_{1.} \dots a_n) = \text{argMax}\, P(a_{1.} \dots a_n|classe) \times P(classe) \quad (3)$$

Assim, Pardo e Nunes (2002) dizem que como o classificador faz a suposição “ingênua” de que todos os atributos da instância que se quer classificar são independentes, o cálculo do termo $P(a_{1.} \dots a_n|classe)$ reduz-se a $P(a_1|classe)$ x … x $P(a_n|classe)$. Logo, a fórmula do classificador seria:

$$\text{argMax}\, P(classe|a_{1.} \dots a_n) = \text{argMax} \prod_i P(a_i|classe) \times P(classe) \quad (4)$$

Carrilho Júnior (2007) diz em seu trabalho que, o classificador de Naive-Bayes é considerado muito eficiente e preciso na rotulação de novas amostras, justificando com a abordagem simples que o classificador trata as características dos exemplos.

Pinto (2005) cita como desvantagem do algoritmo o fato dele assumir que dada uma classe, seus atributos são independentes, hipótese pouco frequente no mundo real. Pardo e Nunes (2002) afirmam que apesar disso o classificador fornece resultados satisfatórios e que quando os atributos são realmente independentes, o classificador fornece a solução ótima.

Outra desvantagem citada por Pinto (2005) é o fato do desempenho do classificador não melhorar muito quando o número de exemplos de treino aumenta.

5 PROCEDIMENTOS METODOLÓGICOS

Esse trabalho foi desenvolvido seguindo as seguintes etapas:

1. Coleta de dados, que consistia em coletar dados gerais e postagens de usuários da rede social Twitter.
2. Identificação e ranqueamento dos usuários, que consistia em aplicar os métodos de identificação e ranqueamento de usuários influentes propostos por Cha *et al*. (2010).
3. Classificação manual das postagens, que consistia na leitura e classificação humana de cada postagem dos usuários pertencentes aos *rankings* gerados.
4. Treinamento e classificação das postagens, que consistia: na transformação dos dados para o formato de entrada do programa *Weka*, na definição da estratégia de treinamento dos dados, no treinamento das amostras escolhidas e na classificação de todas as postagens de cada *ranking*.
5. Categorização dos usuários, que consistia em quantificar o percentual das postagens do usuário pertencia a cada categoria.

Entre as seções 5.1 a 5.5, têm-se mais detalhes de cada etapa.

5.1 Coleta de dados

Para a realização da coleta dos dados, foi desenvolvido um programa na linguagem *PHP* que, por meio de interações com a *REST API* do *Twitter*, armazena em um banco de dados *MySQL* os atributos gerais e dez requisições de postagens de 36 usuários sementes e de seus amigos (pessoas que eles seguem). Os atributos gerais armazenados foram: *user_id*, *name*, *followers_count*, *friends_count*, *screen_name* e *listed_count*. Já os atributos das postagens foram: *id_str_num*, *id_str, in_reply_to_screen_name*, *text*, *in_reply_to_user_id_str*, *in_reply_to_status_id_str*, *lang*, *user_user_id*, *text*, *created_at*, *retweet_count* e *favorite_count*.

Os usuários sementes escolhidos foram selecionados por serem relacionados às categorias política e esporte:

1. AecioNeves
2. AloSenado
3. BocaoPolitica
4. CamaraDeputados
5. CorujaoEsporte
6. ESPNagora
7. Esp_Interativo
8. EstadaoEsporte
9. EstadaoPolitica
10. EulucianaGenro
11. FCGloboEsporte
12. FolhaPolitica
13. GloboEsporteRJ
14. OGloboPolitica
15. OGlobo_Esportes
16. Placar_Esportes
17. RedeTVEsporte
18. SelecaoSporTV
19. SenadoFederal
20. SporTV
21. SuperEsporte
22. TerraEsportesBR
23. UOLEsporte
24. UOLPolitica
25. VEJAEsporte
26. alterosaesp
27. dilmabr
28. espfantastico
29. esportesdc
30. g1politica
31. globoesporte

32. globoesporteSP
33. globoesportecom
34. lancenet
35. politicalivre
36. silva_marina

A coleta ocorreu no período dos dias 25 a 30 de setembro, e armazenou 57.555 usuários e 6.098.079 postagens.

5.2 Identificação e ranqueamento dos usuários influentes

O processo de identificação e o ranqueamento dos usuários influentes ocorreu seguindo os métodos propostos por Cha *et al*. (2010), que consistem em identificar os usuários com mais menções, *retweets* e seguidores, criando-se então três *rankings*.

O *ranking* de menções foi definido por meio da ordenação decrescente do atributo *listed_count* dos usuários. Para definir o *ranking* de *retweets*, foi inicialmente realizada a soma dos atributos *retweet_count* das postagens de cada usuário, não sendo consideras as postagens que já eram *retweets* de outros usuários. Então, foi realizada a ordenação decrescente das somas. Já o *ranking* por seguidores, foi criado a partir da ordenação decrescente do atributo *followers_count* de cada usuário.

Como o objetivo deste trabalho é a categorização das áreas de influência destes usuários baseando-se no conteúdo das postagens, foram removidos dos três *rankings*, os usuários que não possuíam nenhuma postagem identificada pelo *Twitter* como em língua portuguesa, ou seja, com o atributo *lang*=”pt”. Foi também limitado o tamanho destes *rankings* a 100 usuários.

5.3 Classificação manual das postagens

Para a classificação manual das postagens, foram removidas as postagens que não foram identificadas pelo *Twitter* como em língua portuguesa, ou seja, que não possuíam o atributo *lang*=”pt”.

As postagens foram classificadas em quatro classes: *política*, *esporte*, *outros assuntos* e *outro idioma*. A classe *política* representa as postagens que apresentam temas políticos e/ou foram utilizadas em contextos políticos. A classe *esporte*, por sua vez, representa as postagens que apresentam temas esportivos e/ou foram utilizadas em contextos esportivos. A classe *outros assuntos*, representa as postagens que não se enquadram nas duas primeiras categorias.

Apesar de remover previamente as postagens identificadas pelo *Twitter* como em outros idiomas, existiam ainda algumas postagens que não se apresentavam em língua portuguesa. Assim, surgiu-se a necessidade da criação da classe *outro idioma*.

5.4 Treinamento e classificação das postagens

Neste trabalho, optou-se por utilizar um algoritmo de aprendizado supervisionado. O treinamento e a classificação das postagens foram realizados por meio do algoritmo *Naive-Bayes*, com o auxílio da ferramenta *Weka*. Para a realização dos processos de treinamento e classificação, foram eliminadas as postagens que foram classificadas manualmente como *outro idioma*. Após a eliminação destas postagens, restaram: 72 usuários no *ranking* de menções, 59 usuários no *ranking* de *retweets* e 74 usuários no *ranking* de seguidores.

5.4.1 Formatação dos dados

Para a utilização da ferramenta *Weka*, criou-se arquivos de entrada no formato *.arff*. A criação desses arquivos deu-se por meio da execução dos processos descritos nas seções 5.4.1.1 e 5.4.1.2.

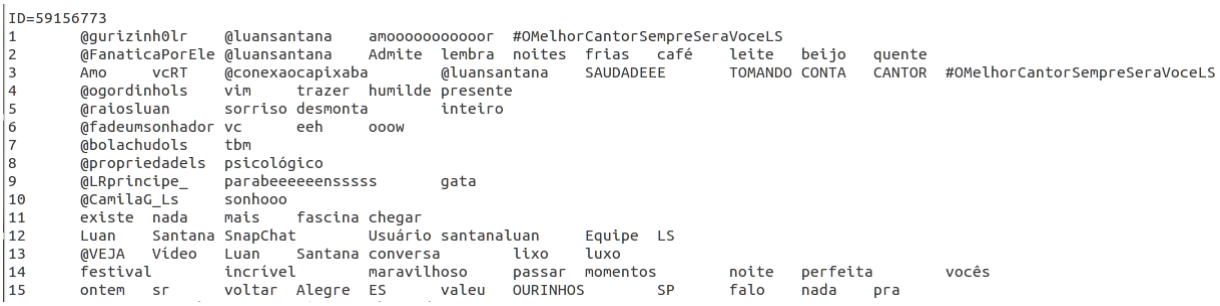
5.4.1.1 Tokenização

O processo de tokenização utilizado, consiste na divisão inicial das postagens por caracteres de espaço, quebra de linha, tabulação e alguns caracteres (",", ":", ";", ")", "(", "]", "[", ".", "$","%", "&", "!", "?", "{", "}", "\"", "\"", "'", "\", "/", "¿","¥", "¡", "|" , "´", "^", "*").

Neste processo, também foram removidas números, *links*, *e-mails* e palavras consideradas *stopwords*. Adicionalmente optou-se por manter os usuários citados nas postagens como *tokens*, partindo do pressuposto de que uma postagem que envolva o usuário @dilmabr possui uma maior probabilidade de ser de assunto político, por exemplo. Mantiveram-se também como *tokens* as *hashtags* contidas nas postagens.

Em seguida, os *tokens* foram armazenados em um documento *.txt* junto ao *user_id* e o número da postagem correspondente, como pode-se ver na Figura 1.

Figura 1: Alguns tokens gerados no ranking de usuários com maiores números de menções



ID=59156773											
1	@gurizinh0lr	@luansantana	amooooooooooor	#OMelhorCantorSempreSeraVoceLS							
2	@FanaticaPorEle	@luansantana	Admite	lembra	noites	frias	café	leite	beijo	quente	
3	Amo	vcRT	@conexaocapixaba	@luansantana	SAUDADEEE		TOMANDO	CONTA	CANTOR	#OMelhorCantorSempreSeraVoceLS	
4	@ogordinhols	vim	trazer	humilde	presente						
5	@raiosluan	sorriso	desmonta	inteiro							
6	@fadeumsonhador	vc	eeh	ooow							
7	@bolachudols	tbm									
8	@propriedadels	psicológico									
9	@LRprincipe_	parabeeeeeensssss		gata							
10	@CamilaG_Ls	sonhooo									
11	existe	nada	mais	fascina	chegar						
12	Luan	Santana	SnapChat	Usuário	santanaluan		Equipe	LS			
13	@VEJA	Video	Luan	Santana	conversa		lixo	luxo			
14	festival		incrível		maravilhoso		passar	momentos		noite	perfeita	vocês
15	ontem	sr	voltar	Alegre	ES		valeu	OURINHOS		SP	falo	nada	pra

Fonte: Produção do próprio autor.

5.4.1.2 Criação dos arquivos de entrada do Weka (.arff)

Os arquivos de entrada da ferramenta *Weka* (.arff) possuíam como atributos: um número identificador do usuário, um número identificador da postagem, todos os *tokens* criados a partir das postagens escolhidas para treinamento do seu *ranking* e a classe a qual a instância pertencia.

Na Figura 2, pode-se ver os atributos iniciais e na Figura 3 podemos ver os atributos finais de um dos arquivos *.arff* gerado para uma das classificações realizadas.

Figura 2: Atributos iniciais de um arquivo .arff gerado

```
@relation mencoesEntrada
@attribute idusuario string
@attribute numpostagem numeric
@attribute a0 { 0, 1 }
@attribute a1 { 0, 1 }
@attribute a2 { 0, 1 }
@attribute a3 { 0, 1 }
@attribute a4 { 0, 1 }
```

Fonte: Produção do próprio autor.

Figura 3: Atributos finais de um arquivo .arff gerado

```
@attribute a6418 { 0, 1 }
@attribute a6419 { 0, 1 }
@attribute a6420 { 0, 1 }
@attribute a6421 { 0, 1 }
@attribute a6422 { 0, 1 }
@attribute classe { 'Esporte', 'Política', 'Outros assuntos' }
```

Fonte: Produção do próprio autor.

As instâncias do arquivo correspondiam às postagens escolhidas para treinamento e/ou validação que não pertenciam à classe *outro idioma*. Cada instância possuía o *user_id* do seu autor, um identificador criado para sua postagem, e uma sequência de 0's e 1's que representavam em ordem se a postagem possuía aquele atributo ou *token* (0-não possui, 1- possui). Por fim, a instância contia a classe correspondente. Na Figura 4 têm-se exemplos de instâncias criadas.

Figura 4: Instâncias para classificação

```
@data
'155659213',2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
'155659213',3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
'155659213',4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

Fonte: Produção do próprio autor.

5.4.2 Execução do treinamento e classificação das postagens

Durante esta etapa, foram realizados dez treinamentos e classificações por *ranking*. As amostras treinadas foram selecionadas aleatoriamente. O número de amostras de treinamento para cada classe foi definido pelo percentual de 70% do total de amostras da classe com menos elementos. O processo de classificação foi feito com todas as amostras dos rankings.

5.5 Categorização dos usuários

Para a categorização dos usuários, foi criado um programa na linguagem Java que, por meio de interações com o algoritmo *Naive-Bayes* implementado pela ferramenta *Weka*, obtém a classificação de cada instância (postagem). Com as postagens já classificadas, o algoritmo calcula o percentual de postagens do usuário que se enquadra em cada categoria (*esporte*, *política* e *outros assuntos*). Para a realização desta etapa, foram selecionados os resultados obtidos após os treinamentos e classificações que obtiveram o maior e o menor percentual de instâncias classificadas corretamente de cada *ranking.*

6 RESULTADOS OBTIDOS

6.1 Identificação e ranqueamento dos usuários influentes

Após a finalização da etapa de identificação e ranqueamento dos usuários influentes, foram gerados três *rankings*: o primeiro com os usuários que possuíam o maior número de menções, o qual se nomeou de *ranking* Menções; o segundo com os usuários que possuíam o maior número de *retweets*, o qual se nomeou de *ranking Retweets*; e o último com os usuários que possuíam o maior número de seguidores, o qual se nomeou de *ranking* Seguidores.

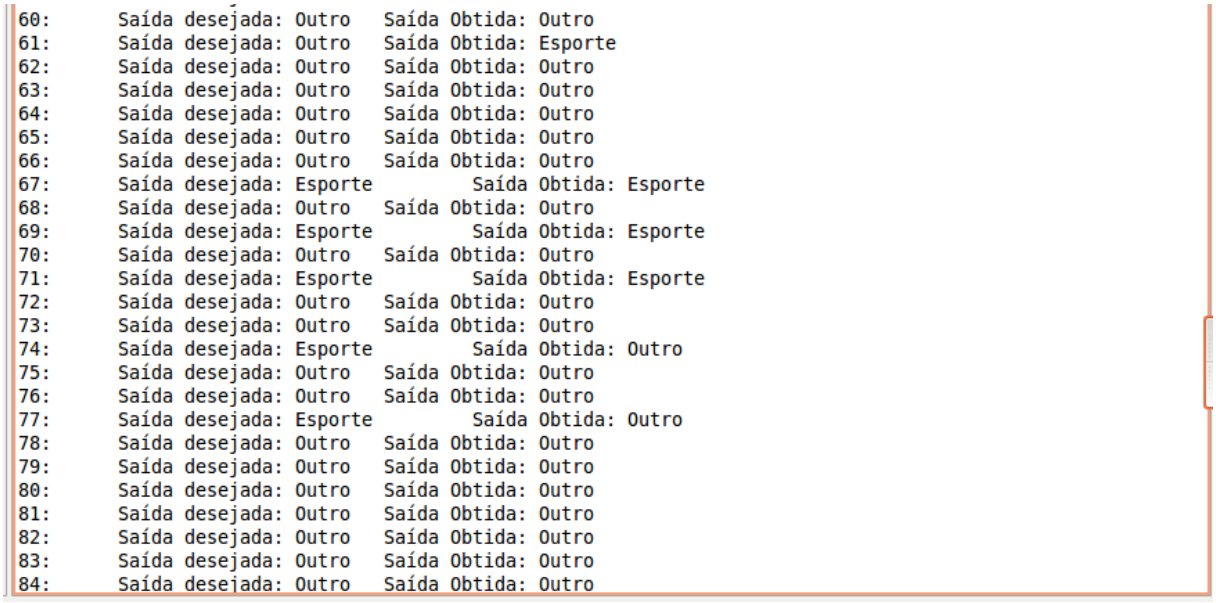
Cada um destes rankings possuía 100 usuários, destes: 7 pertenciam simultaneamente aos três *rankings*, 19 pertenciam simultaneamente aos rankings Menções e *Retweets*, 29 pertenciam simultaneamente aos rankings Menções e Seguidores e 25 pertenciam simultaneamente aos rankings *Retweets* e Seguidores.

6.2 Classificação das postagens

O processo de classificação das postagens gerou dois tipos de resultados: o resultado geral do treinamento e classificação, com matriz de confusão, dados sobre a acurácia do modelo, entre outros do modelo, e o resultado obtido por instância.

Na Figura 5, têm-se um exemplo de resultado obtido por instância de um dos processos de classificação realizado para o *ranking* Seguidores.

Figura 5: Resultados por instância do processo de classificação do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do *ranking* Seguidores



Fonte: Produção do próprio autor.

Entre as seções 6.2.1 e 6.2.3 têm-se os resultados gerais dos processos de classificação realizados para cada um dos *rankings*.

6.2.1 Ranking Menções

O primeiro *ranking* a ter os dados treinados e classificados foi o de Menções. Após realizar os dez treinamentos e classificações, obteve-se o percentual médio de instâncias classificadas corretamente de 76.67562%. O desvio padrão obtido foi de 0.68196%. O maior percentual de instâncias classificadas corretamente obtido foi de 78.1806%, enquanto o menor foi de 75.8868%. O percentual de instâncias classificadas corretamente em cada treinamento e classificação realizado é apresentado na tabela 1.

Tabela 1: Percentual de instâncias classificadas corretamente em cada treinamento e classificação do *ranking* Menções

Número do treinamento e classificação:	Instâncias corretamente classificadas (%):
1	78.1806

Número do treinamento e classificação:	Instâncias corretamente classificadas (%):
2	76.9945
3	75.9539
4	76.9721
5	76.7372
6	75.8868
7	76.3455
8	76.9498
9	76.6924
10	76.0434

Fonte: Produção do próprio autor.

6.2.1.1 Maior percentual de instâncias classificadas corretamente

O treinamento contou com 6424 atributos e 1776 instâncias, sendo 592 pertencentes a cada classe. A classificação foi realizada com 8937 instâncias. Este grupo obteve 78.1806% das instâncias classificadas corretamente. Na classificação das postagens pertencentes à classe *esporte*, obteve-se 58.6785% de acertos. Para a classe *política*, obteve-se 59.2857% de acertos. E por fim, a classe *outros assuntos* obteve a taxa de 83.2133% de acertos. A matriz de confusão da classificação se encontra na tabela 2.

Tabela 2: Matriz de confusão do treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente do ranking Menções

a	b	c	
595	49	370	**a=Esporte**
46	498	296	**b=Política**
619	570	5894	**c=Outros assuntos**

Fonte: Produção do próprio autor.

Na tabela 3 têm-se dados sobre a acurácia da classificação. Pode-se observar que a precisão do algoritmo para a classe *esporte* foi de 0.472, para a

classe *política* foi de 0.446, e para a classe outros *assuntos* foi de 0.898. A média ponderada entre a precisão das classes foi de 0.808.

Tabela 3: Acurácia do treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente do ranking Menções

Classe:	Taxa de verdadeiros positivos (TP)	Taxa de falsos positivos (FP)	Precisão	Revocação (Recall)	F-Measure	Área ROC
Esporte	0.587	0.084	0.472	0.587	0.523	0.821
Política	0.593	0.076	0.446	0.593	0.509	0.829
Outros assuntos	0.832	0.359	0.898	0.832	0.864	0.799
Média Ponderada	0.782	0.301	0.808	0.782	0.792	0.804

Fonte: Produção do próprio autor.

6.2.1.2 Menor percentual de instâncias classificadas corretamente

O treinamento contou com 6381 atributos e 1776 instâncias, sendo 592 pertencentes a cada classe. A classificação foi realizada com 8937 instâncias. Este grupo obteve 75.8868% das instâncias classificadas corretamente. Na classificação das postagens pertencentes à classe *esporte*, obteve-se 61.4398% de acertos. Para a classe *política*, obteve-se 62.7381% de acertos. E por fim, a classe *outros assuntos* obteve a taxa de 79.5143% de acertos. A matriz de confusão da classificação se encontra na tabela 4.

Tabela 4: Matriz de confusão do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do ranking Menções

a	b	c	
623	64	327	**a=Esporte**
62	527	251	**b=Política**

a	b	c	
826	625	5632	**c=Outros assuntos**

Fonte: Produção do próprio autor.

Na tabela 5 têm-se dados sobre a acurácia da classificação. Pode-se observar que a precisão do algoritmo para a classe *esporte* foi de 0.412, para a classe *política* foi de 0.433, e para a classe *outros assuntos* foi de 0.907. A média ponderada entre a precisão das classes foi de 0.806.

Tabela 5: Acurácia do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do ranking Menções

Classe:	Taxa de verdadeiros positivos (TP)	Taxa de falsos positivos (FP)	Precisão	Revocação (Recall)	F-Measure	Área ROC
Esporte	0.614	0.112	0.412	0.614	0.493	0.81
Política	0.627	0.085	0.433	0.627	0.513	0.839
Outros assuntos	0.795	0.312	0.907	0.795	0.847	0.804
Média Ponderada	0.759	0.268	0.806	0.759	0.776	0.808

Fonte: Produção do próprio autor.

6.2.1.3 Análise compartiva entre o maior e o menor percentual de instâncias classificadas corretamente

Um comparativo entre os treinamentos e classificações que obtiveram o maior e o menor percentual de instâncias classificadas corretamente é apresentado na tabela 6.

Tabela 6: Comparativo entre os treinamentos e classificações que obtiveram o maior e o menor percentual de instâncias classificadas corretamente no ranking Menções

Dados:	Maior percentual de instâncias classificadas corretamente:	Menor percentual de instâncias classificadas corretamente:
Número de atributos da amostra	6424	6381
Número de instâncias utilizadas no treinamento	1776	1776
Número de instâncias classificadas	8937	8937
Instâncias classificadas corretamentes (%)	78.1806	75.8868
Instâncias classificadas incorretamentes (%)	21.8194	24.1132
Estatística Kappa	0.4413	0.4222
Média do erro absoluto	0.1995	0.2168
Erro quadrático médio	0.3496	0.3635
Erro absoluto relativo (%)	44.8839	48.7707
Raíz do erro quadrado relativo (%)	74.1622	77.1106

Fonte: Produção do próprio autor.

Na tabela 7, têm-se os erros e acertos de ambos os treinamentos e classificações por classes. Para a classe *esporte*, o treinamento e classificação com o menor percentual de instâncias classificadas corretamente obteve o maior percentual de acertos, com 61.4398%. Para a classe *política*, este mesmo processo de treinamento e classificação obteve o maior percentual de acertos, com 62.7381%. Na classe *outros assuntos*, o treinamento e classificação com maior percentual de instâncias classificadas corretamente obteve o maior percentual de acertos, com 83.2133%.

Tabela 7: Percentual de erros e acertos por classe da classificação das postagens do *ranking* Menções

Treinamento e classificação:	Esporte		Política		Outros assuntos	
	Erros (%)	Acertos (%)	Erros (%)	Acertos (%)	Erros (%)	Acertos (%)
Maior percentual de instâncias classificadas corretamente	41.3215	58.6785	40.7143	59.2857	16.7867	83.2133
Menor percentual de instâncias classificadas corretamente	38.5602	61.4398	37.2619	62.7381	20.4857	79.5143

Fonte: Produção do próprio autor.

6.2.2 Ranking Retweets

O segundo *ranking* a ter os dados treinados e classificados foi o de *Retweets*. Após realizar os dez treinamentos e classificações, obteve-se o percentual médio de instâncias classificadas corretamente de 73.11837%. O desvio padrão obtido foi de 1.19034%. O maior percentual de instâncias classificadas corretamente obtido foi 74.9621%, enquanto o menor foi de 71.5984%. O percentual de instâncias classificadas corretamente em cada treinamento e classificação realizado é apresentado na tabela 8.

Tabela 8: Percentual de instâncias classificadas corretamente em cada treinamento e classificação do *ranking* Retweets

Número do treinamento e classificação:	Instâncias corretamente classificadas (%):
1	73.4952
2	74.2792
3	71.6237
4	73.3687
5	72.7112
6	72.8376
7	71.5984

Número do treinamento e classificação:	Instâncias corretamente classificadas (%):
8	74.3804
9	71.9272
10	74.9621

Fonte: Produção do próprio autor.

6.2.2.1

6.2.2.2 Maior percentual de instâncias classificadas corretamente

O treinamento contou com 2562 atributos e 516 instâncias, sendo 172 pertencentes a cada classe. A classificação foi realizada com 3954 instâncias. Este grupo obteve 74.9621% das instâncias classificadas corretamente. Na classificação das postagens pertencentes à classe *esporte*, obteve-se 53.8690% de acertos. Para a classe *política*, obteve-se 61.3821% de acertos. E por fim, a classe *outros assuntos* obteve a taxa de 84.0741% de acertos. A matriz de confusão da classificação se encontra na tabela 9.

Tabela 9: Matriz de confusão do treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente do ranking Reweets

a	b	c	
543	18	447	**a=Esporte**
10	151	85	**b=Política**
285	145	2270	**c=Outros assuntos**

Fonte: Produção do próprio autor.

Na tabela 10 têm-se dados sobre a acurácia da classificação. Pode-se observar que a precisão do algoritmo para a classe *esporte* foi de 0.648, para a classe *política* foi de 0.481, e para a classe *outros assuntos* foi de 0.81. A média ponderada entre a precisão das classes foi de 0.748.

Tabela 10: Acurácia do treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente do ranking Retweets

Classe:	Taxa de verdadeiros positivos (TP)	Taxa de falsos positivos (FP)	Precisão	Revocação (Recall)	F-Measure	Área ROC
Esporte	0.539	0.1	0.648	0.539	0.588	0.784
Política	0.614	0.044	0.481	0.614	0.539	0.872
Outros assuntos	0.841	0.424	0.81	0.841	0.825	0.754
Média Ponderada	0.75	0.318	0.748	0.75	0.747	0.769

Fonte: Produção do próprio autor.

6.2.2.3

6.2.2.4 Menor percentual de instâncias classificadas corretamente

O treinamento contou com 2525 atributos e 516 instâncias, sendo 172 pertencentes a cada classe. A classificação foi realizada com 3954 instâncias. Este grupo obteve 71.5984% das instâncias classificadas corretamente. Na classificação das postagens pertencentes à classe *esporte*, obteve-se 66.8651% de acertos. Para a classe *política*, obteve-se 63.8211% de acertos. E por fim, a classe *outros assuntos* obteve a taxa de 74.0741% de acertos. A matriz de confusão da classificação se encontra na tabela 11.

Tabela 11: Matriz de confusão do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do ranking Retweets

a	b	c	
674	14	320	**a=Esporte**
25	157	64	**b=Política**
586	114	2000	**c=Outros assuntos**

Fonte: Produção do próprio autor.

Na tabela 12 têm-se dados sobre a acurácia da classificação. Pode-se observar que a precisão do algoritmo para a classe *esporte* foi de 0.525, para a

classe *política* foi de 0.551, e para a classe *outros assuntos* foi de 0.839. A média ponderada entre a precisão das classes foi de 0.741.

Tabela 12: Acurácia do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do ranking Retweets

Classe:	Taxa de verdadeiros positivos (TP)	Taxa de falsos positivos (FP)	Precisão	Revocação (Recall)	F-Measure	Área ROC
Esporte	0.669	0.207	0.525	0.669	0.588	0.793
Política	0.638	0.035	0.551	0.638	0.591	0.887
Outros assuntos	0.741	0.306	0.839	0.741	0.787	0.768
Média Ponderada	0.716	0.264	0.741	0.716	0.724	0.782

Fonte: Produção do próprio autor.

6.2.2.5 Análise compartiva entre o maior e o menor percentual de instâncias classificadas corretamente

Um comparativo entre os treinamentos e classificações que obtiveram o maior e o menor percentual de instâncias classificadas corretamente é apresentado na tabela 13.

.

Tabela 13: Comparativo entre os treinamentos e classificações que obtiveram o maior e o menor percentual de instâncias classificadas corretamente no *ranking Retweets*

Dados:	Maior percentual de instâncias classificadas corretamente	Menor percentual de instâncias classificadas corretamente
Número de atributos da amostra	2562	2525
Número de instâncias utilizadas no treinamento	516	516
Número de instâncias classificadas	3954	3954

Dados:	Maior percentual de instâncias classificadas corretamente	Menor percentual de instâncias classificadas corretamente
Instâncias classificadas corretamentes (%)	74.9621	71.5984
Instâncias classificadas incorretamentes (%)	25.0379	28.4016
Estatística Kappa	0.4523	0.433
Média do erro absoluto	0.2289	0.2522
Erro quadrático médio	0.3645	0.3823
Erro absoluto relativo (%)	51.5111	56.7541
Raíz do erro quadrado relativo (%)	77.3317	81.0956

Fonte: Produção do próprio autor.

Na tabela 14, têm-se os erros e acertos de ambos os treinamentos e classificações por classes. Para a classe *esporte*, o treinamento e classificação com o menor percentual de instâncias classificadas corretamente obteve o maior percentual de acertos, com 66.8651%. Para a classe *política*, este mesmo processo de treinamento e classificação obteve o maior percentual de acertos, com 63.8211%. Na classe *outros assuntos*, o treinamento e classificação com maior percentual de instâncias classificadas corretamente obteve o maior percentual de acertos, com 84.0741%.

Tabela 14: Percentual de erros e acertos por classe da classificação das postagens do *ranking Retweets*

Treinamento e Classificação:	Esporte		Política		Outros assuntos	
	Erros (%)	Acertos (%)	Erros (%)	Acertos (%)	Erros (%)	Acertos (%)
Maior percentual de instâncias classificadas corretamente	46.1310	53.8690	38.6179	61.3821	15.9259	84.0741

Treinamento e Classificação:	Esporte		Política		Outros assuntos	
	Erros (%)	Acertos (%)	Erros (%)	Acertos (%)	Erros (%)	Acertos (%)
Menor percentual de instâncias classificadas corretamente	33.1349	66.8651	36.1789	63.8211	25.9259	74.0741

Fonte: Produção do próprio autor.

6.2.3 Ranking Seguidores

O último *ranking* a ter os dados treinados e classificados foi o de Seguidores. Após realizar os dez treinamentos e classificações, obteve-se o percentual médio de instâncias classificadas corretamente de 80.00103%. O desvio padrão obtido foi de 0.84887%. O maior percentual de instâncias classificadas corretamente foi de 80.9936% das instâncias classificadas corretamente, enquanto o menor foi de 78.1488%. O percentual de instâncias classificadas corretamente em cada treinamento e classificação realizado é apresentado na tabela 15.

Tabela 15: Percentual de instâncias classificadas corretamente em cada treinamento e classificação do *ranking* Seguidores

Número do treinamento e classificação:	Instâncias corretamente classificadas (%):
1	78.1488
2	80.1278
3	80.9215
4	79.4063
5	80.334
6	80.9936
7	80.3958
8	80.2618
9	80.1278

Número do treinamento e classificação:	Instâncias corretamente classificadas (%):
10	79.2929

Fonte: Produção do próprio autor.

6.2.3.1 Maior percentual de instâncias classificadas corretamente

O treinamento contou com 4766 atributos e 1287 instâncias, sendo 429 pertencentes a cada classe. A classificação foi realizada com 9702 instâncias. Este grupo obteve 80.9936% das instâncias classificadas corretamente. Na classificação das postagens pertencentes à classe *esporte*, obteve-se 69.5910% de acertos. Para a classe *política*, obteve-se 54.0230% de acertos. E por fim, a classe *outros assuntos* obteve a taxa de 85.4428% de acertos. A matriz de confusão da classificação se encontra na tabela 16.

Tabela 16: Matriz de confusão do treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente do ranking Seguidores

a	b	c	
1055	78	383	**a=Esporte**
58	329	222	**b=Política**
510	593	6474	**c=Outros assuntos**

Fonte: Produção do próprio autor.

Na tabela 17 têm-se dados sobre a acurácia da classificação. Pode-se observar que a precisão do algoritmo para a classe *esporte* foi de 0.65, para a classe *política* foi de 0.329, e para a classe *outros assuntos* foi de 0.915. A média ponderada entre a precisão das classes foi de 0.836.

Tabela 17: Acurácia do treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente do *ranking* Seguidores

Classe:	Taxa de verdadeiros positivos (TP)	Taxa de falsos positivos (FP)	Precisão	Revocação (Recall)	F-Measure	Área ROC
Esporte	0.696	0.069	0.65	0.696	0.672	0.87
Política	0.54	0.074	0.329	0.54	0.409	0.813
Outros assuntos	0.854	0.285	0.915	0.854	0.883	0.849
Média Ponderada	0.81	0.238	0.836	0.81	0.821	0.85

Fonte: Produção do próprio autor.

6.2.3.2 Menor percentual de instâncias classificadas corretamente

O treinamento contou com 4724 atributos e 1287 instâncias, sendo 429 pertencentes a cada classe. A classificação foi realizada com 9702 instâncias. Este grupo obteve 78.1488% das instâncias classificadas corretamente. Na classificação das postagens pertencentes à classe *esporte*, obteve-se 70.7784% de acertos. Para a classe *política*, obteve-se 55.1724% de acertos. E por fim, a classe *outros assuntos* obteve a taxa de 81.4570% de acertos. A matriz de confusão da classificação se encontra na tabela 18.

Tabela 18: Matriz de confusão do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do *ranking* Seguidores

a	b	c	
1073	97	346	**a=Esporte**
66	337	206	**b=Política**
629	776	6172	**c=Outros assuntos**

Fonte: Produção do próprio autor.

Na tabela 19 têm-se dados sobre a acurácia da classificação. Pode-se observar que a precisão do algoritmo para a classe *esporte* foi de 0.607, para a classe *política* foi de 0.279, e para a classe *outros assuntos* foi de 0.918. A média ponderada entre a precisão das classes foi de 0.829.

Tabela 19: Acurácia do treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente do *ranking* Seguidores

Classe:	Taxa de verdadeiros positivos (TP)	Taxa de falsos positivos (FP)	Precisão	Revocação (Recall)	F-Measure	Área ROC
Esporte	0.708	0.085	0.607	0.708	0.653	0.865
Política	0.553	0.096	0.279	0.553	0.371	0.801
Outros assuntos	0.815	0.26	0.918	0.815	0.863	0.842
Média Ponderada	0.781	0.222	0.829	0.781	0.799	0.843

Fonte: Produção do próprio autor.

6.2.3.3 Análise compartiva entre o maior e o menor percentual de instâncias classificadas corretamente

Um comparativo entre os treinamentos e classificações que obtiveram o maior e o menor percentual de instâncias classificadas corretamente é apresentado na tabela 20.

Tabela 20: Comparativo entre os treinamentos e classificações que obtiveram o maior e o menor percentual de instâncias classificadas corretamente no *ranking* Seguidores

Dados:	Maior percentual de instâncias classificadas corretamente	Menor percentual de instâncias classificadas corretamente
Número de atributos da amostra	4766	4724
Número de instâncias utilizadas no treinamento	1287	1287

Dados:	Maior percentual de instâncias classificadas corretamente	Menor percentual de instâncias classificadas corretamente
Número de instâncias classificadas	9702	9702
Instâncias classificadas corretamentes (%)	80.9936	78.1488
Instâncias classificadas incorretamentes (%)	19.0064	21.8512
Estatística Kappa	0.5219	0.4827
Média do erro absoluto	0.1900	0.2051
Erro quadrático médio	0.3272	0.3421
Erro absoluto relativo (%)	42.7607	46.1429
Raíz do erro quadrado relativo (%)	69.3991	72.5762

Fonte: Produção do próprio autor.

Na tabela 21, têm-se os erros e acertos de ambos os treinamentos e classificações por classes. Para classe *esporte*, o treinamento e classificação com o menor percentual de instâncias classificadas corretamente obteve o maior percentual de acertos, com 70.7784%. Para a classe *política*, este mesmo treinamento e classificação obteve o maior percentual de acertos, com 55.1724%. Na classe *outros assuntos*, o treinamento e classificação com maior percentual de instâncias classificadas corretamente obteve o maior percentual de acertos, com 85.4428%.

Tabela 21: Percentual de erros e acertos por classe da classificação das postagens do *ranking* Seguidores

Treinamento e classificação:	Esporte		Política		Outros assuntos	
	Erros (%)	Acertos (%)	Erros (%)	Acertos (%)	Erros (%)	Acertos (%)

Maior percentual de instâncias classificadas corretamente	30.4090%	69.5910%	45.9770%	54.0230%	14.5572%	85.4428%
Menor percentual de instâncias classificadas corretamente	29.2216%	70.7784%	44.8276%	55.1724%	18.5430%	81.4570%

Fonte: Produção do próprio autor.

6.3 Categorização dos usuários

Após a etapa de classificação das postagens, foi realizada a categorização dos usuários de cada *ranking*, baseando-se nos resultados obtidos nos treinamentos e classificações que obtiveram o maior e o menor percentual de instâncias classificadas corretamente. Neste processo foi tomado as classes como categorias e encontrado o percentual de pertencimento do usuário a cada uma delas.

Um exemplo de saída do processo de categorização está na Figura 6, onde se têm os percentuais obtidos na categorização do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do *ranking* Seguidores, enquanto na Figura 7 temos os percentuais reais desses mesmos usuários.

Figura 6: Percentuais obtidos na categorização do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do *ranking* Seguidores

```
Valores Obtidos
ID: 155659213   Total de Postagens: 5   Política: 0,00000       Esporte: 0,40000        Outros: 0,60000
ID: 60865434    Total de Postagens: 67  Política: 0,01493       Esporte: 0,14925        Outros: 0,83582
ID: 158487331   Total de Postagens: 143 Política: 0,00699       Esporte: 0,11189        Outros: 0,88112
ID: 14872237    Total de Postagens: 2   Política: 0,00000       Esporte: 0,50000        Outros: 0,50000
ID: 96951800    Total de Postagens: 5   Política: 0,00000       Esporte: 0,80000        Outros: 0,20000
ID: 14502789    Total de Postagens: 176 Política: 0,02273       Esporte: 0,07386        Outros: 0,90341
ID: 29479000    Total de Postagens: 131 Política: 0,00000       Esporte: 0,44275        Outros: 0,55725
ID: 48332360    Total de Postagens: 158 Política: 0,01899       Esporte: 0,04430        Outros: 0,93671
ID: 14213711    Total de Postagens: 334 Política: 0,38024       Esporte: 0,02395        Outros: 0,59581
ID: 90836187    Total de Postagens: 3   Política: 0,00000       Esporte: 1,00000        Outros: 0,00000
ID: 43023158    Total de Postagens: 171 Política: 0,02924       Esporte: 0,09942        Outros: 0,87135
ID: 224223563   Total de Postagens: 2   Política: 0,00000       Esporte: 0,50000        Outros: 0,50000
ID: 5520952     Total de Postagens: 55  Política: 0,05455       Esporte: 0,12727        Outros: 0,81818
ID: 34171224    Total de Postagens: 168 Política: 0,00595       Esporte: 0,00595        Outros: 0,98810
```

Fonte: Produção do próprio autor

Figura 7: Percentuais reais na categorização do treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do ranking Seguidores

```
Valores Reais
ID: 155659213   Total de Postagens: 5   Política: 0,00000       Esporte: 0,80000        Outros: 0,20000
ID: 60865434    Total de Postagens: 67  Política: 0,01493       Esporte: 0,19403        Outros: 0,79104
ID: 158487331   Total de Postagens: 143 Política: 0,00000       Esporte: 0,18182        Outros: 0,81818
ID: 14872237    Total de Postagens: 2   Política: 0,00000       Esporte: 0,50000        Outros: 0,50000
ID: 96951800    Total de Postagens: 5   Política: 0,00000       Esporte: 0,20000        Outros: 0,80000
ID: 14502789    Total de Postagens: 176 Política: 0,00000       Esporte: 0,00568        Outros: 0,99432
ID: 29479000    Total de Postagens: 131 Política: 0,00000       Esporte: 0,57252        Outros: 0,42748
ID: 48332360    Total de Postagens: 158 Política: 0,00000       Esporte: 0,00000        Outros: 1,00000
ID: 14213711    Total de Postagens: 334 Política: 0,20958       Esporte: 0,00299        Outros: 0,78743
ID: 90836187    Total de Postagens: 3   Política: 0,00000       Esporte: 0,66667        Outros: 0,33333
ID: 43023158    Total de Postagens: 171 Política: 0,00000       Esporte: 0,05848        Outros: 0,94152
ID: 224223563   Total de Postagens: 2   Política: 0,00000       Esporte: 0,50000        Outros: 0,50000
ID: 5520952     Total de Postagens: 55  Política: 0,00000       Esporte: 0,00000        Outros: 1,00000
ID: 34171224    Total de Postagens: 168 Política: 0,00000       Esporte: 0,01786        Outros: 0,98214
```

Fonte: Produção do próprio autor

Entre as seções 6.3.1 e 6.3.3 têm-se a comparação gráfica do percentual obtido para cada categoria com o percentual real, para cada um dos treinamentos e classificações. Têm-se também um diagrama de dispersão criado para a verificação de correlação entre os valores obtidos com os valores reais.

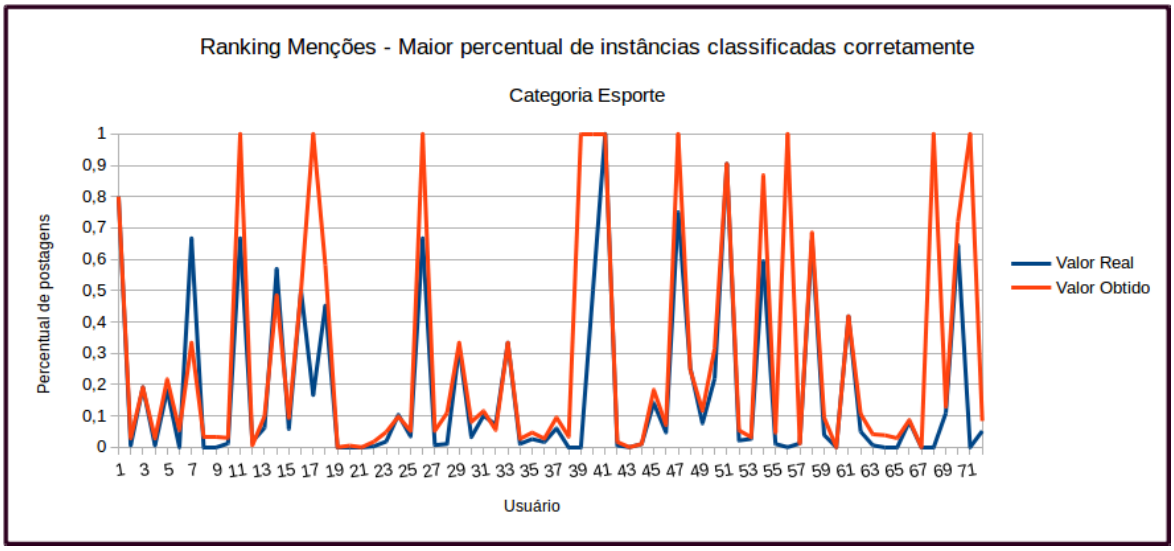
6.3.1 Ranking Menções

6.3.1.1 Maior percentual de instâncias classificadas corretamente

O primeiro processo de categorização foi realizado com o resultado obtido do treinamento e classificação do *ranking* Menções que apresentou o maior percentual de instâncias classificadas corretamente.

O percentual de pertencimento dos usuários a categoria *esporte* é apresentado no Gráfico 1. Os usuários que apresentaram maiores diferenças entre os valores reais e os valores obtidos foram: 7, 11, 17, 26, 39, 40, 47, 54, 56, 68 e 71. A maior diferença apresentada entre estes valores foi de 1.0000.

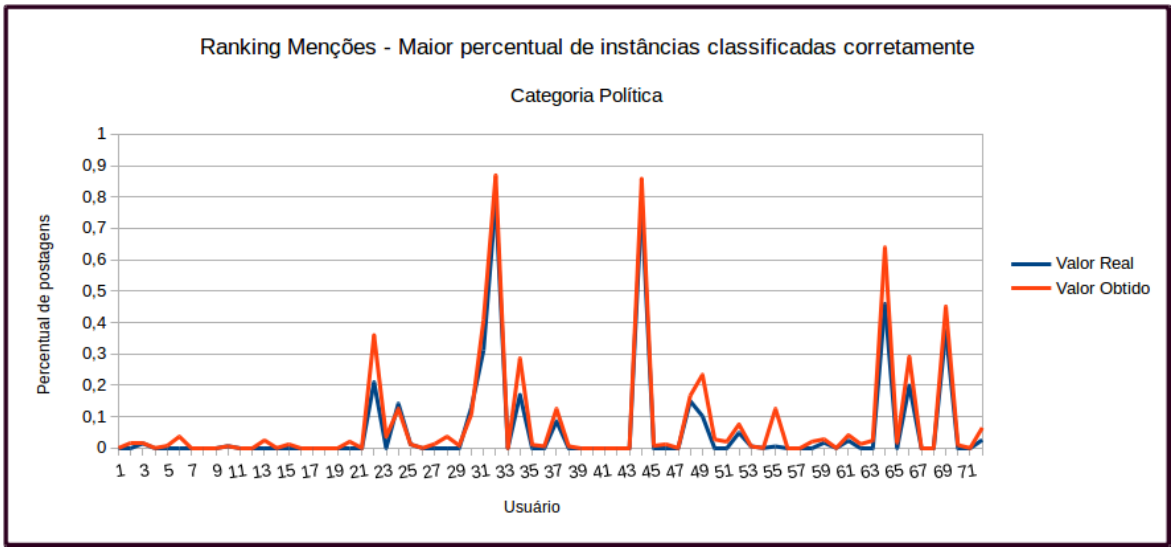
Gráfico 1: Percentual de pertencimento a categoria esporte dos usuários do *ranking* Menções, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

O percentual de pertencimento dos usuários a categoria *política* é apresentado no Gráfico 2. Pode-se perceber que na categoria *política* têm-se uma melhor aproximação entre os valores reais e os valores obtidos, apresentando uma diferença significativa apenas nos usuários: 22, 34, 49, 55, 64 e 66. A maior diferença apresentada foi de 0.18032.

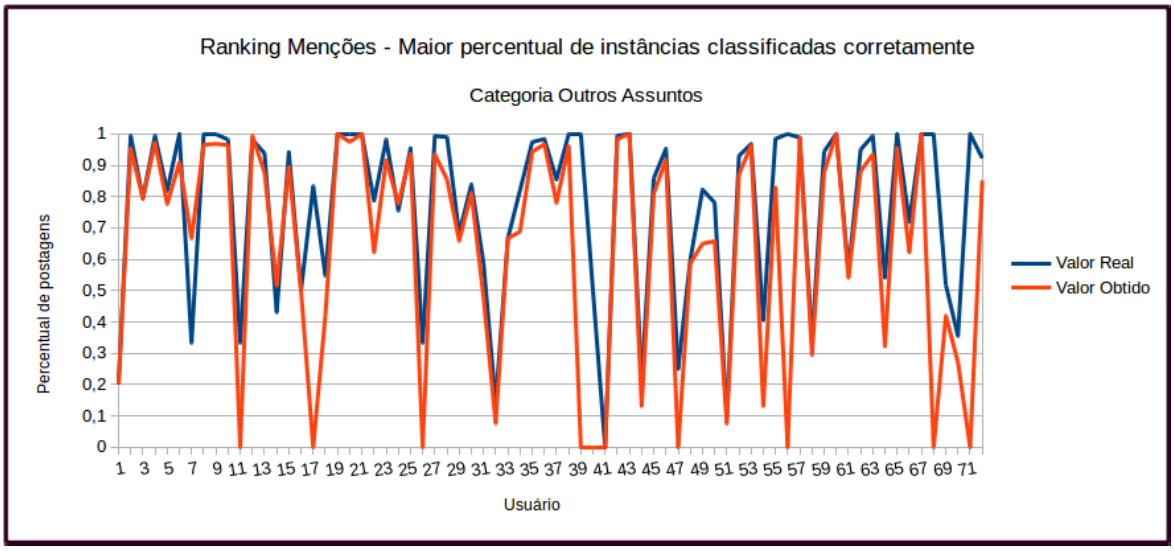
Gráfico 2: Percentual de pertencimento a categoria *política* dos usuários do *ranking* Menções, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

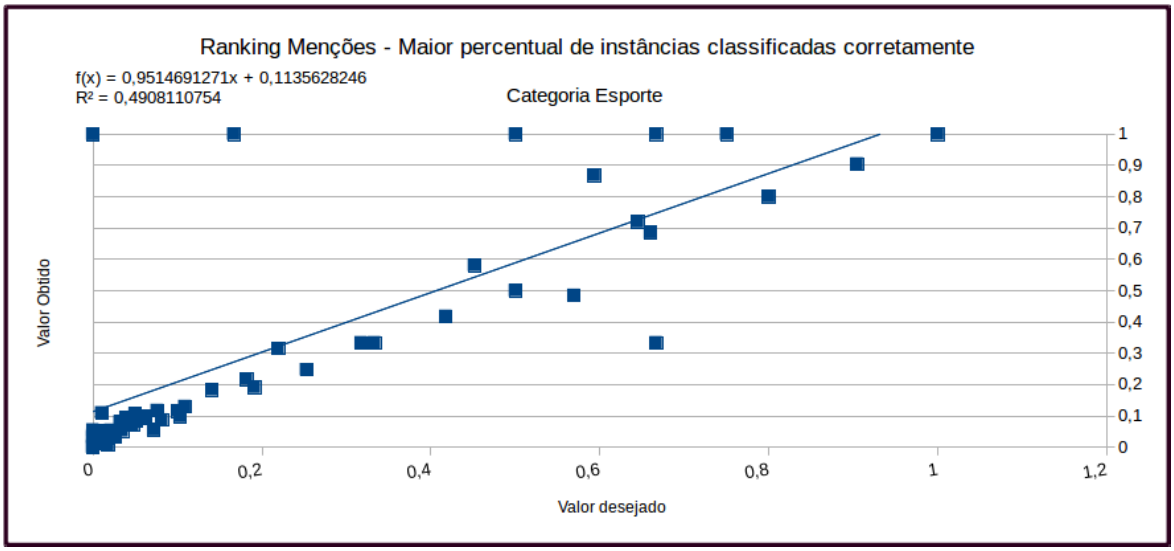
O percentual de pertencimento dos usuários a categoria *outros assuntos* é apresentado no Gráfico 3. Os usuários que apresentaram maiores diferenças entre os valores reais e os valores obtidos foram: 7, 11, 17, 22, 26, 39, 40, 47, 49, 50, 54, 56, 64, 68 e 71. A maior diferença apresentada entre estes valores foi de 1.0000.

Gráfico 3: Percentual de pertencimento a categoria *outros assuntos* dos usuários do *ranking* Menções, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

Gráfico 4: Correlação entre o valor obtido e o valor real no percentual de pertencimento dos usuários à categoria esporte

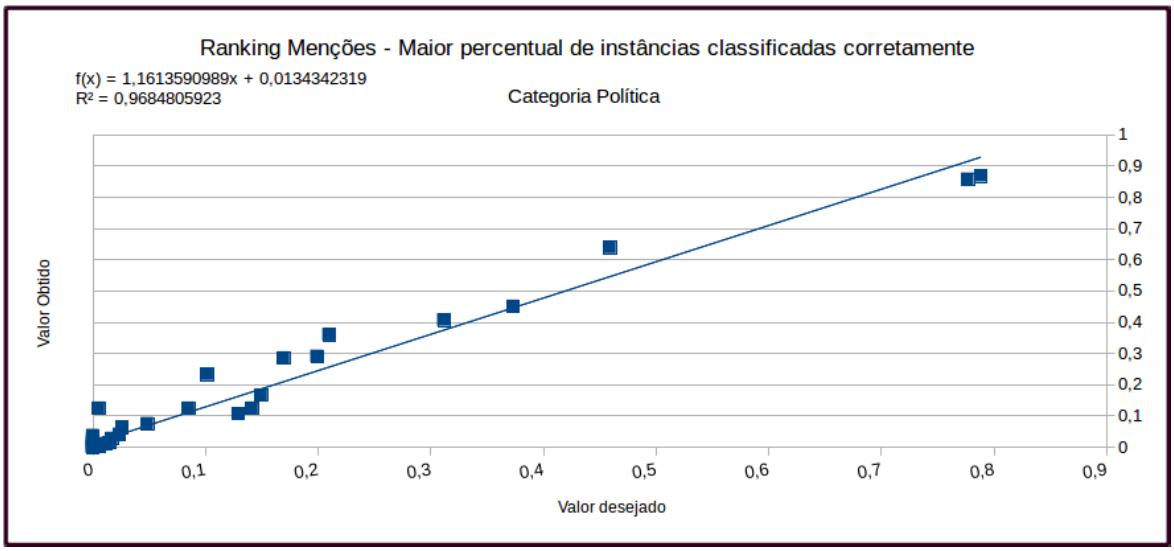


Fonte: Produção do próprio autor.

No Gráfico 4 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *esporte*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.9514691271x + 0.1135628246$ e possui o coeficiente de determinação $R^2 = 0.4908110754$, ou seja, 49.08110754% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

No Gráfico 5 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *política*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 1.1613590989x + 0.0134342319$ e possui o coeficiente de determinação $R^2 = 0.9684805923$. Logo, 96.84805923% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

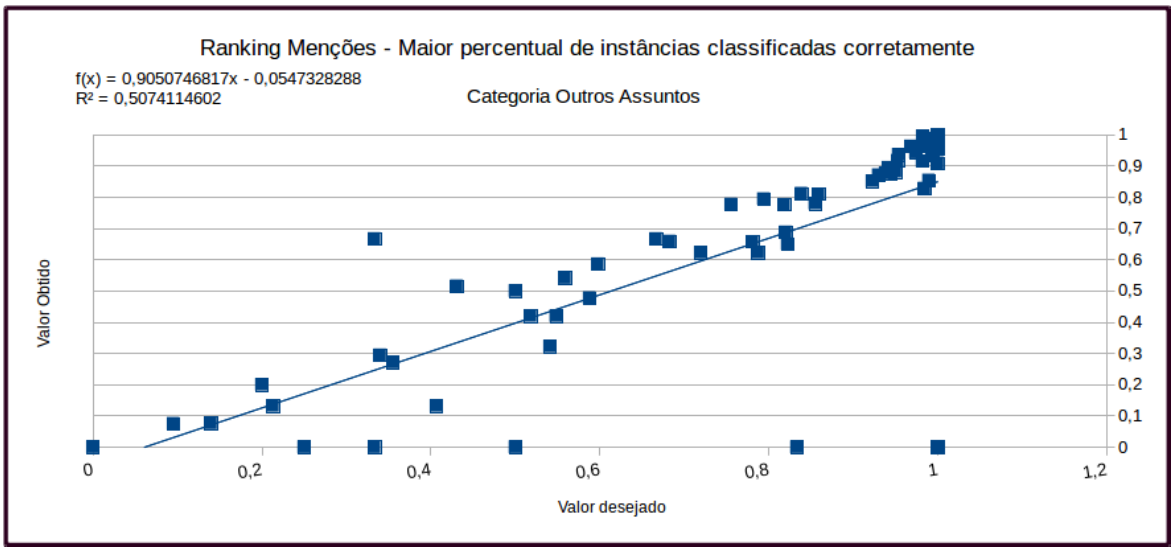
Gráfico 5: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *política*



Fonte: Produção do próprio autor.

No Gráfico 6 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *outros assuntos*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.9050746817x + 0.0547328288$ e possui o coeficiente de determinação $R^2 = 0.5074114602$. Então, 50.74114602% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

Gráfico 6: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *outros assuntos*



Fonte: Produção do próprio autor.

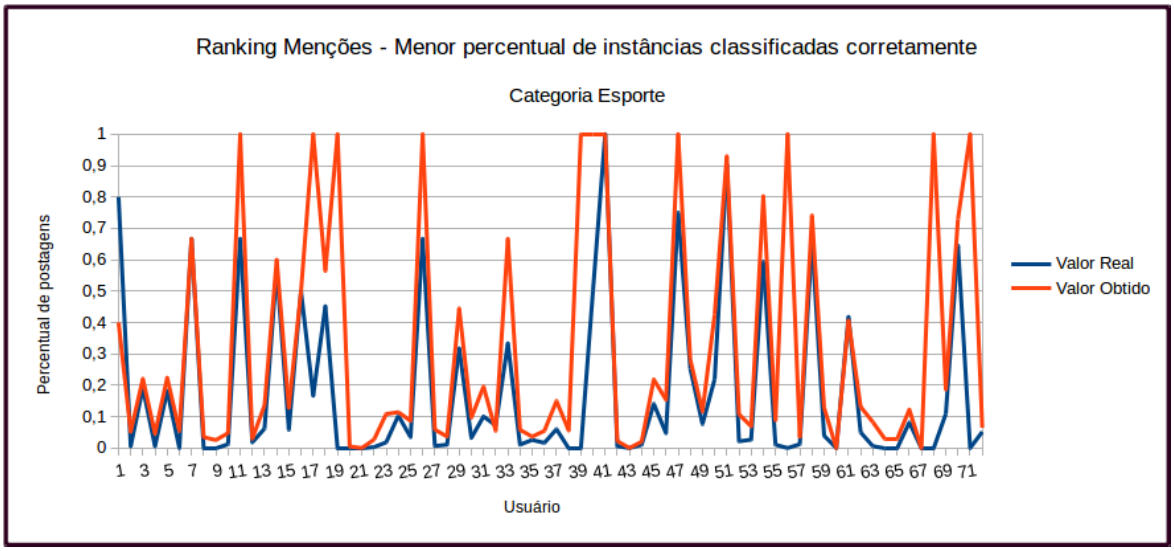
De todas as categorias analisadas, a categoria *política* foi a que apresentou a menor diferença entre o percentual de postagens real e o obtido por usuário e a maior correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários.

6.3.1.2 Menor percentual de instâncias classificadas corretamente

O segundo processo de categorização foi realizado com o resultado obtido do treinamento e classificação do *ranking* Menções que apresentou o menor percentual de instâncias classificadas corretamente.

O percentual de pertencimento dos usuários a categoria *esporte* é apresentado no Gráfico 7. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 11, 17, 18, 19, 26, 29, 31, 32, 37, 39, 40, 45, 47, 54, 56, 68 e 71. A maior diferença apresentada entre estes valores foi de 1.0000.

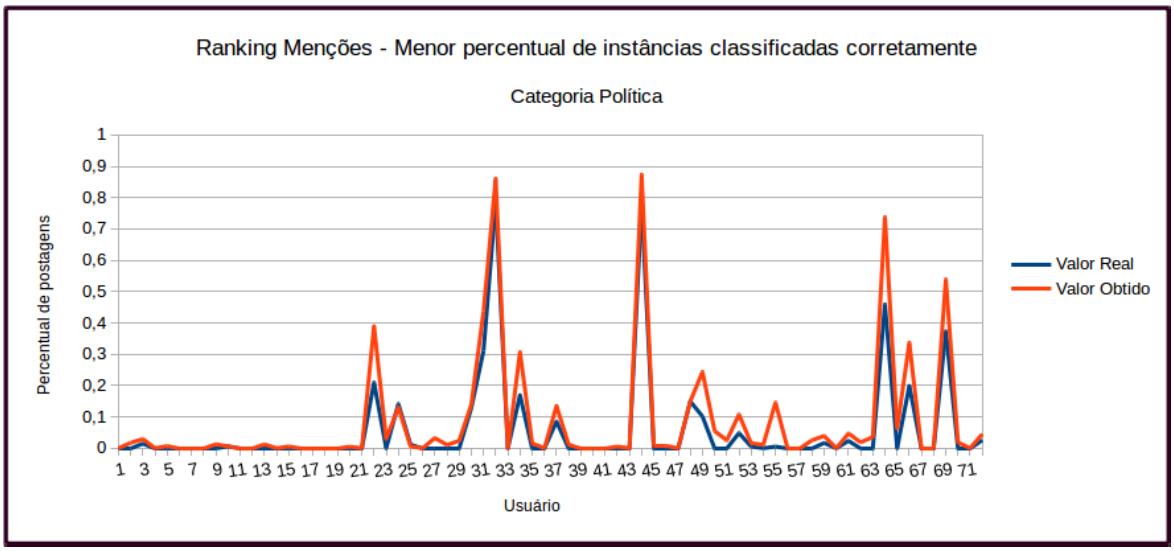
Gráfico 7: Percentual de pertencimento a categoria esporte dos usuários do *ranking* Menções, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

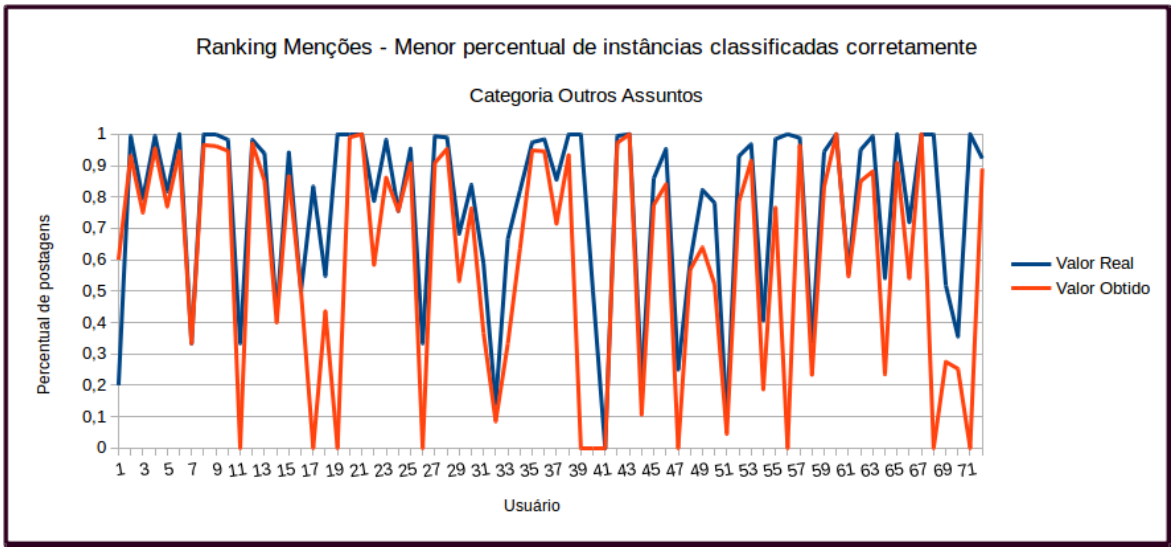
O percentual de pertencimento dos usuários a categoria *política* é apresentado no Gráfico 8. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 22, 34, 49, 50, 55, 64, 66 e 69. A maior diferença apresentada entre estes valores foi de 0.27868.

Gráfico 8: Percentual de pertencimento a categoria *política* dos usuários do *ranking* Menções, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor

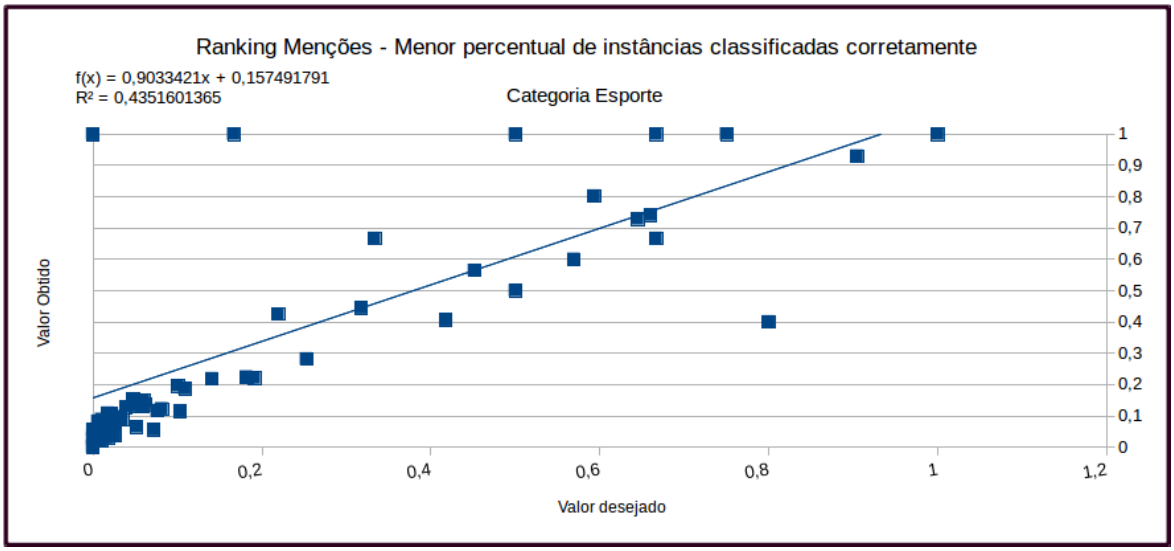
Gráfico 9: Percentual de pertencimento a categoria *outros assuntos* dos usuários do *ranking* Menções, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

O percentual de pertencimento dos usuários a categoria *outros assuntos* é apresentado no Gráfico 9. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 11, 17, 18, 19, 22, 39, 40, 46, 47, 55, 56, 62, 63, 64, 68, 69, 70 e 71. A maior diferença apresentada entre estes valores foi de 1.00000.

Gráfico 10: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *esporte*

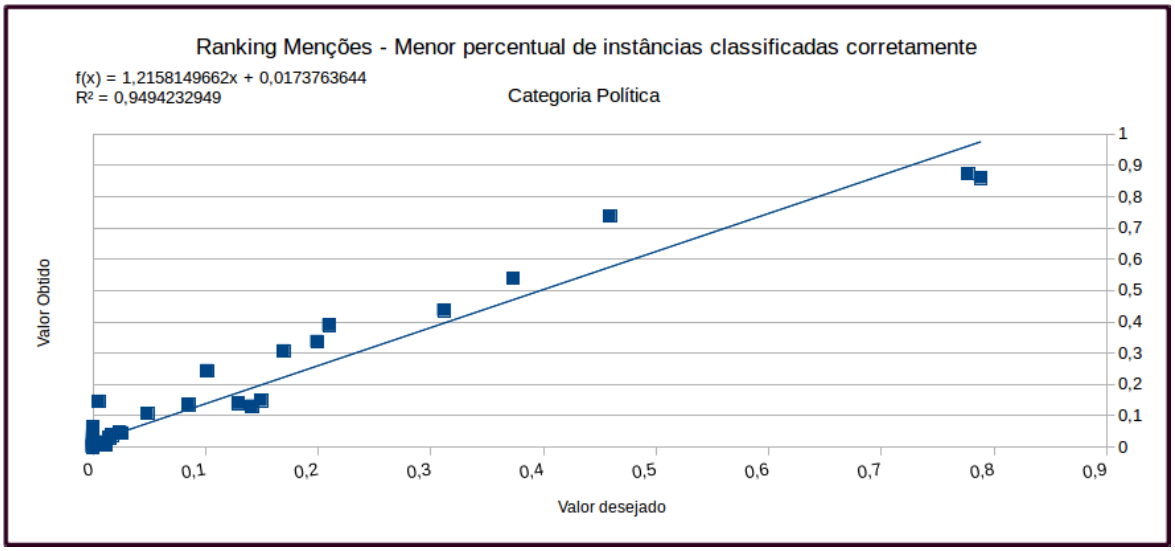


Fonte: Produção do próprio autor.

No Gráfico 10 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *esporte*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.9033421x + 0.157491791$ e possui o coeficiente de determinação $R^2 = 0.4351601365$. Conclui-se então que 43.51601365% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

No Gráfico 11 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *política*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 1.2158149662x + 0.0173763644$ e possui o coeficiente de determinação $R^2 = 0.9494232949$. Conclui-se então que 94.94232949% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

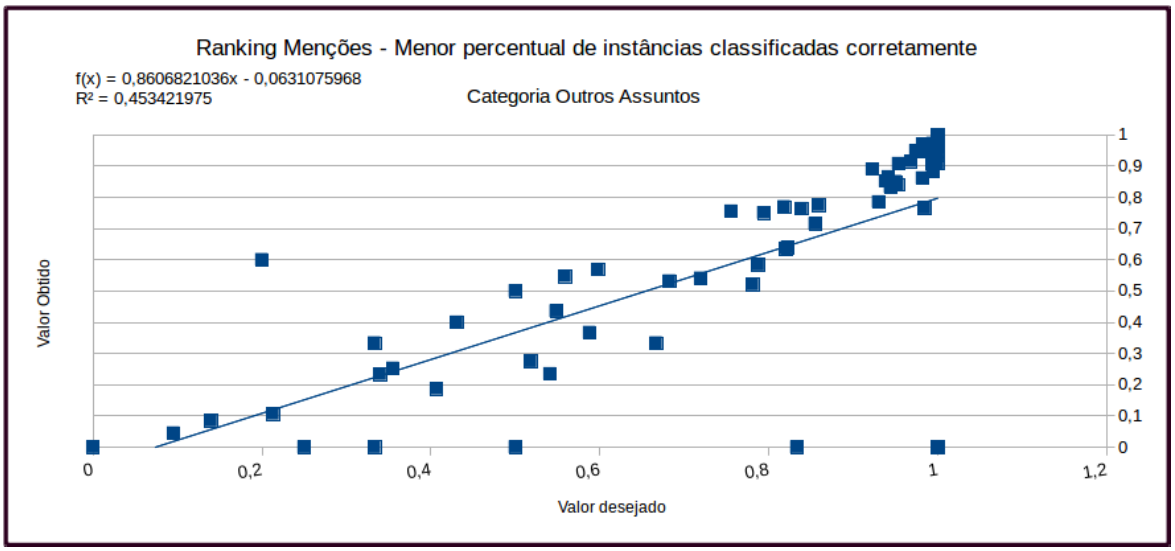
Gráfico 11: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *política*



Fonte: Produção do próprio autor.

No Gráfico 12 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *outros assuntos*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.8606821036x + 0.0631075968$ e possui o coeficiente de determinação $R^2 = 0.453421975$. Conclui-se então que 45.3421975% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

Gráfico 12: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *outros assuntos*



Fonte: Produção do próprio autor.

Assim como na categorização realizada com o resultado do treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente, a categoria *política* foi a que apresentou a menor diferença entre o percentual de postagens real e o obtido por usuário e a maior correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários.

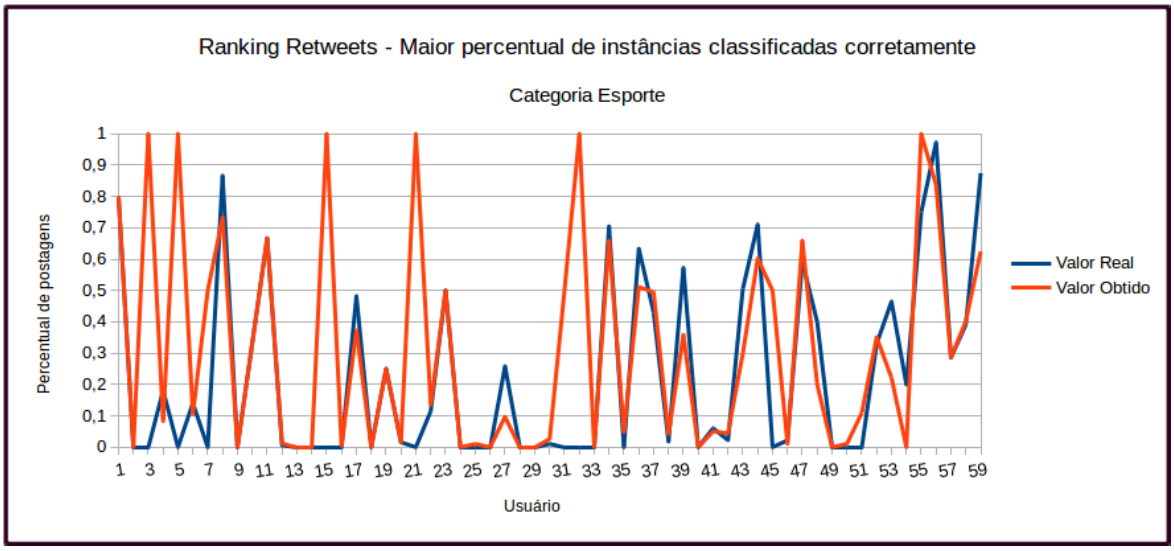
6.3.2 Ranking Retweets

6.3.2.1 Maior percentual de instâncias classificadas corretamente

O terceiro processo de categorização foi realizado com o resultado obtido do treinamento e classificação do *ranking Retweets* que apresentou o maior percentual de instâncias classificadas corretamente.

O percentual de pertencimento dos usuários a categoria *esporte* é apresentado no Gráfico 13. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 3, 5, 7, 15, 21, 27, 31, 32, 39, 45, 53 e 55. A maior diferença apresentada entre estes valores foi de 1.00000.

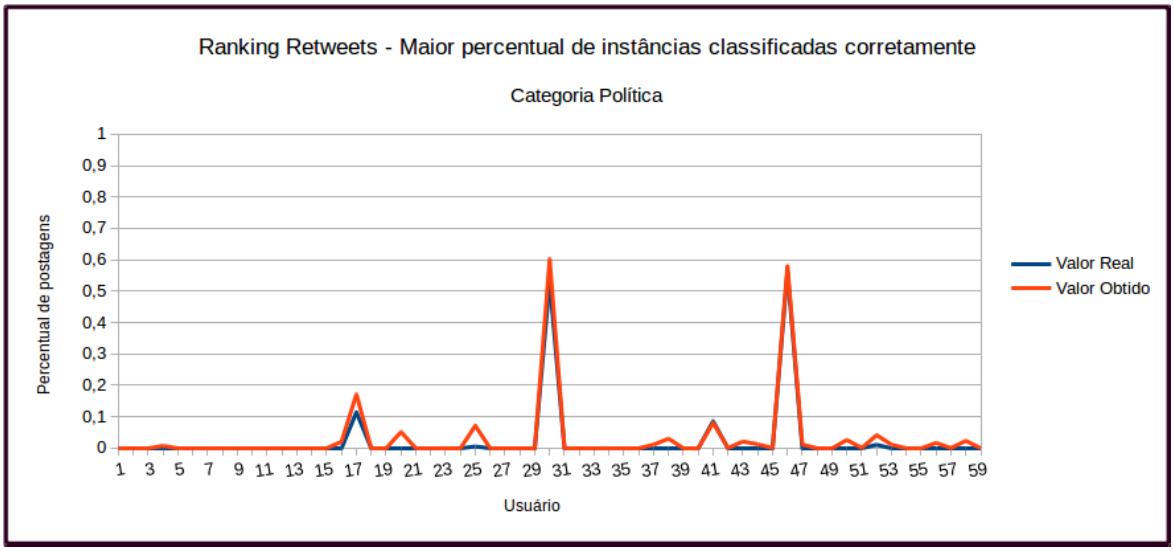
Gráfico 13: Percentual de pertencimento a categoria *esporte* dos usuários do *ranking* Retweets, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

O percentual de pertencimento dos usuários a categoria *política* é apresentado no Gráfico 14. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 20, 25, 38, 43, 50, 52. A maior diferença apresentada entre estes valores foi de 0.07329.

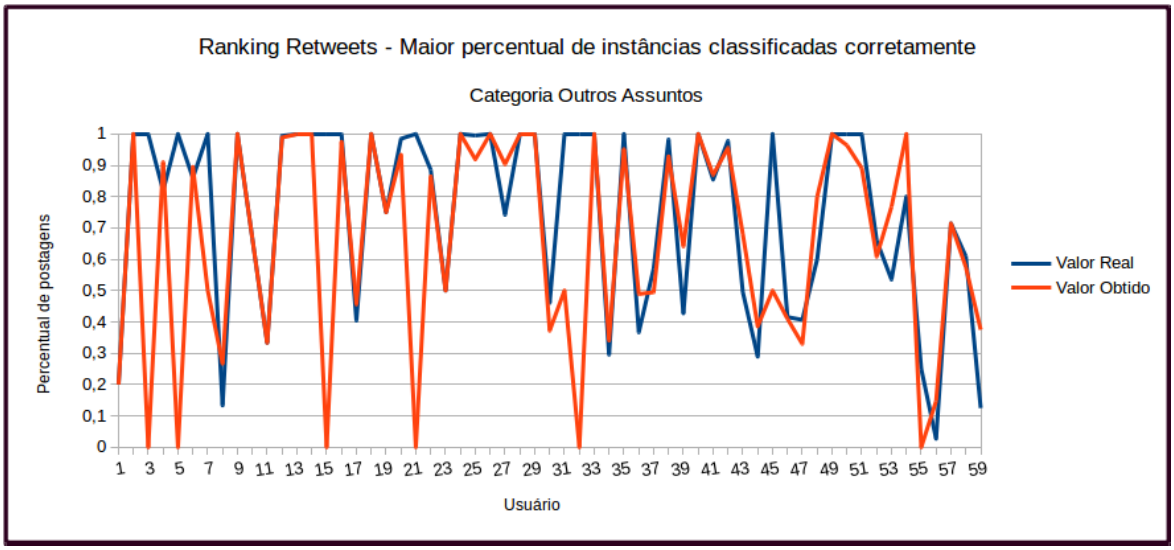
Gráfico 14: Percentual de pertencimento a categoria *política* dos usuários do *ranking* Retweets, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

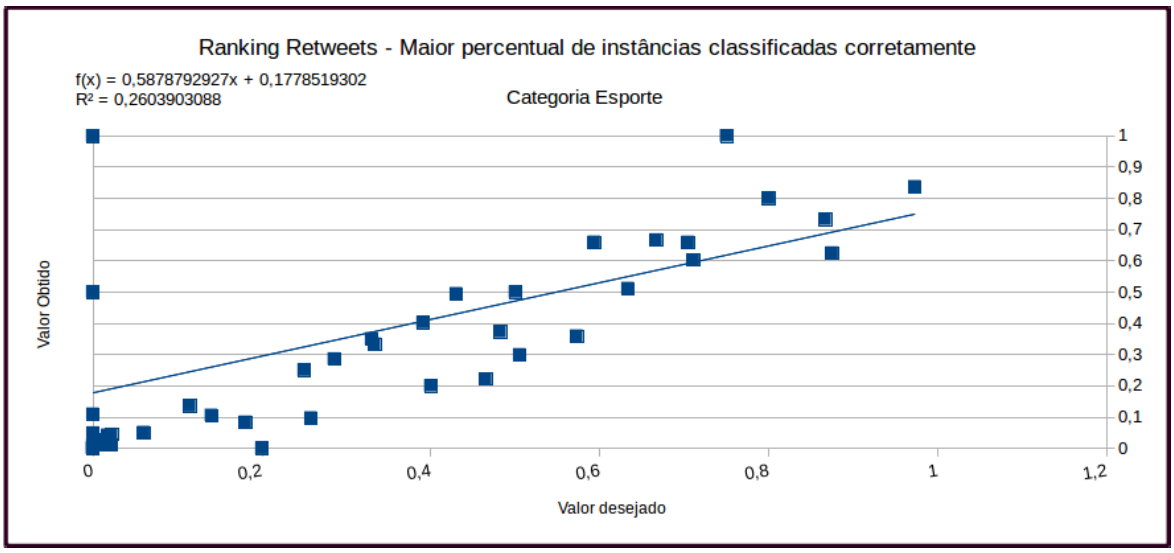
O percentual de pertencimento dos usuários a categoria *outros assuntos* é apresentado no Gráfico 15. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 3, 5, 7, 15, 21, 25, 27,31, 32, 45, 50, 53, 54 e 55. A maior diferença apresentada entre estes valores foi de 1.00000.

Gráfico 15: Percentual de pertencimento a categoria *outros assuntos* dos usuários do *ranking* Retweets, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

Gráfico 16: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *esporte*

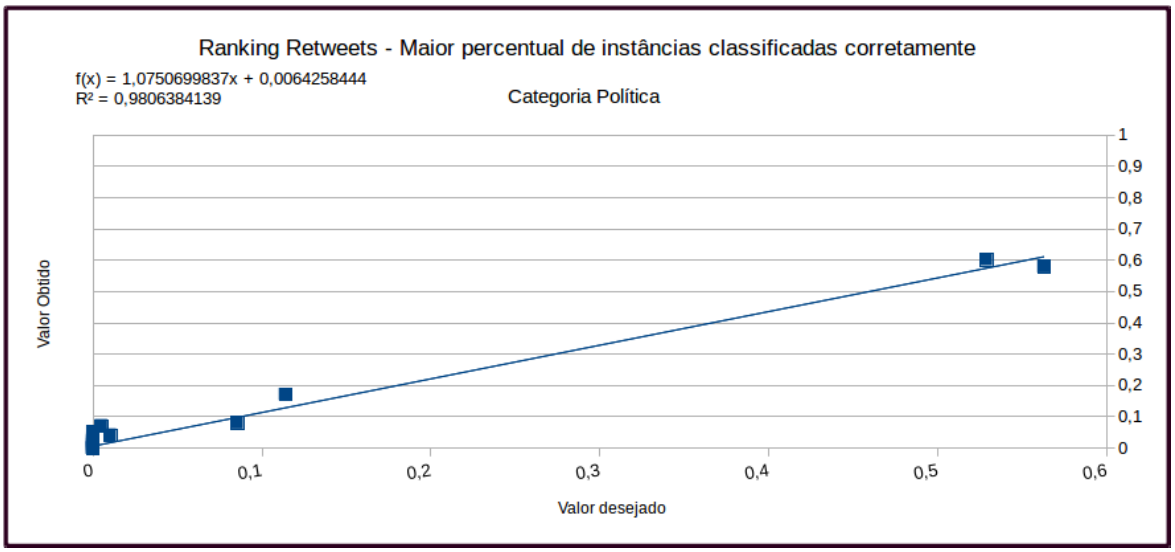


Fonte: Produção do próprio autor.

No Gráfico 16 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *esporte*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.5878792927x + 0.1778519302$ e possui o coeficiente de determinação $R^2 = 0.2603903088$. Então, 26.03903088% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

No Gráfico 17 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *política*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 1.0750699837x + 0.0064258444$ e possui o coeficiente de determinação $R^2 = 0.9806384139$. Logo, 98.06384139% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

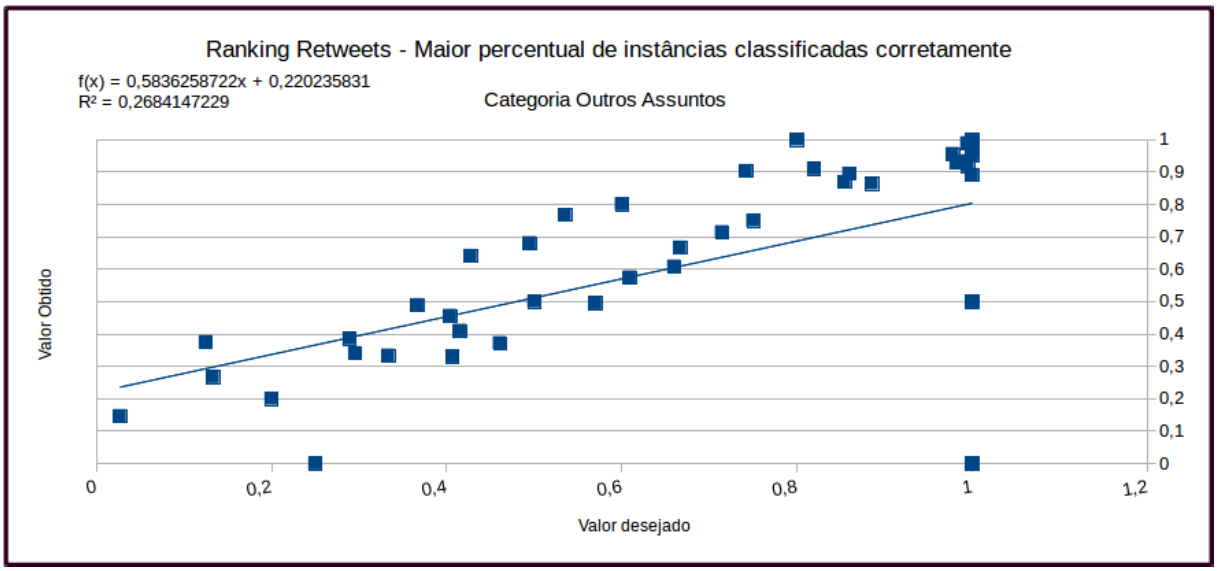
Gráfico 17: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *política*



Fonte: Produção do próprio autor.

No Gráfico 18 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *política*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.5836258722x + 0.220235831$ e possui o coeficiente de determinação $R^2 = 0.2684147229$. Logo, 26.84147229% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

Gráfico 18: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *outros assuntos*



Fonte: Produção do próprio autor.

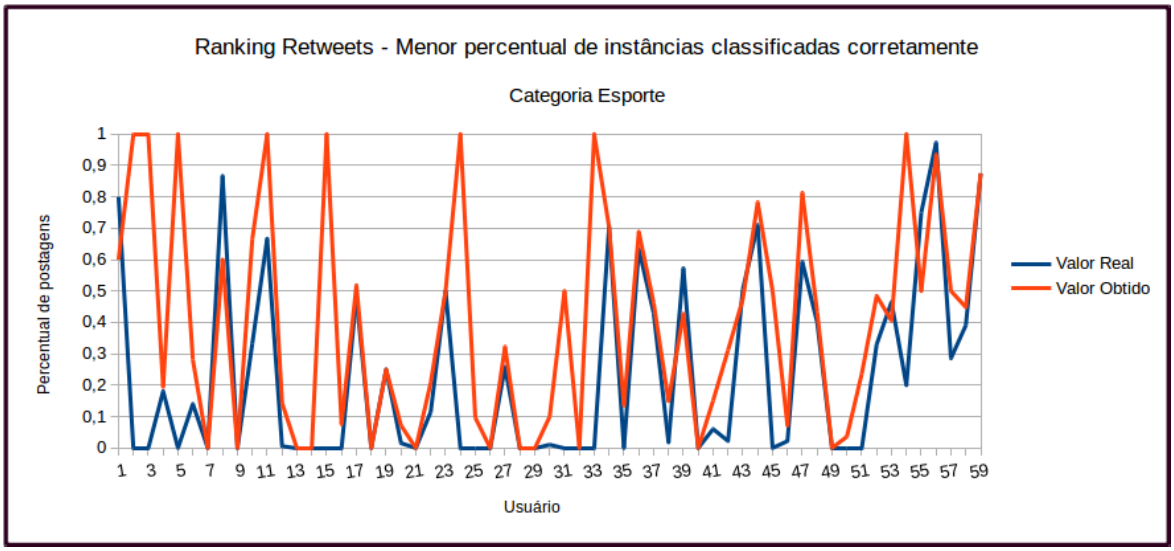
Novamente a categoria *política* foi a que apresentou a menor diferença entre o percentual de postagens real e o obtido por usuário e a maior correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários. Em contrapartida, a correlação encontrada para as categorias *esporte* e *outros assuntos* foram inferiores a 30%.

6.3.2.2 Menor percentual de instâncias classificadas corretamente

O quarto processo de categorização foi realizado com o resultado obtido do treinamento e classificação do *ranking Retweets* que apresentou o menor percentual de instâncias classificadas corretamente.

O percentual de pertencimento dos usuários a categoria *esporte* é apresentado no Gráfico 19. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 1, 2, 3, 5, 6, 8, 10, 11, 12, 15, 24, 25, 30, 31, 33, 42, 47, 50, 51, 54 e 57. A maior diferença apresentada entre estes valores foi de 1.00000.

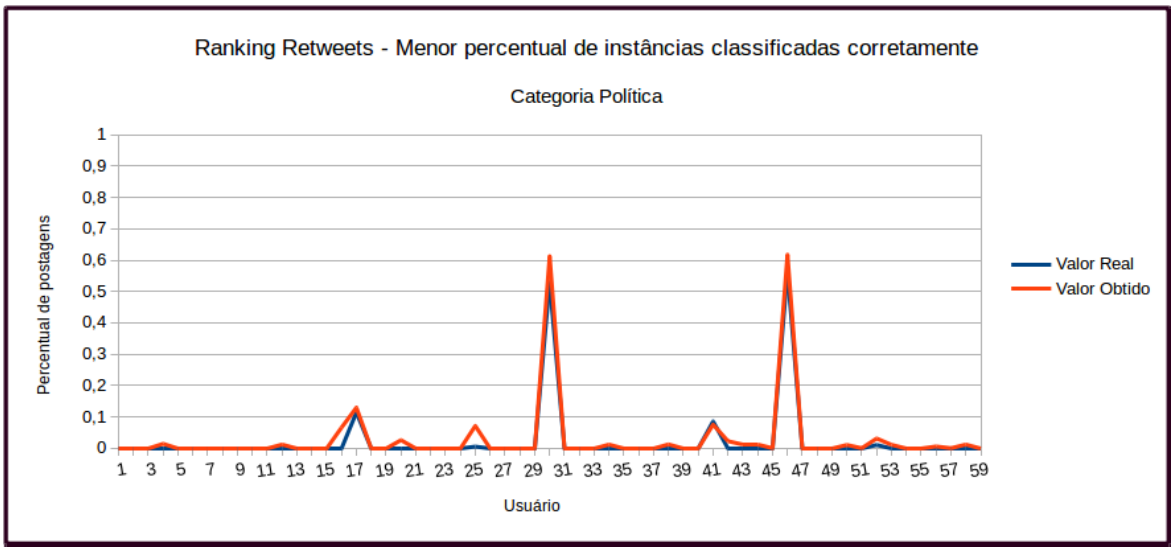
Gráfico 19: Percentual de pertencimento a categoria *esporte* dos usuários do *ranking* Retweets, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

O percentual de pertencimento dos usuários a categoria *política* é apresentado no Gráfico 20. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 16, 20 e 25. A maior diferença apresentada entre estes valores foi de 0.08377.

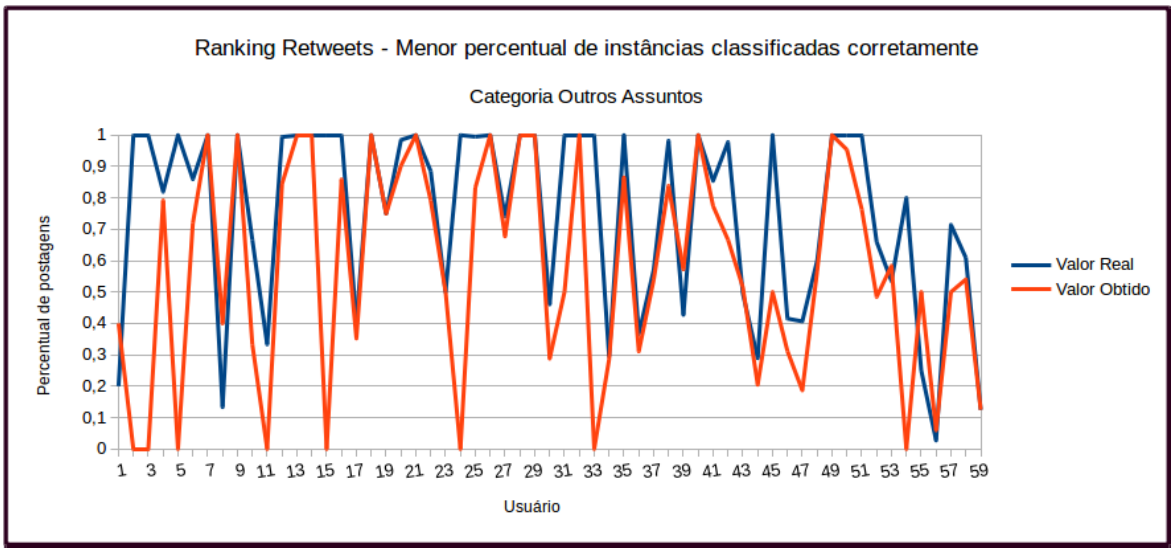
Gráfico 20: Percentual de pertencimento a categoria *política* dos usuários do *ranking* Retweets, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

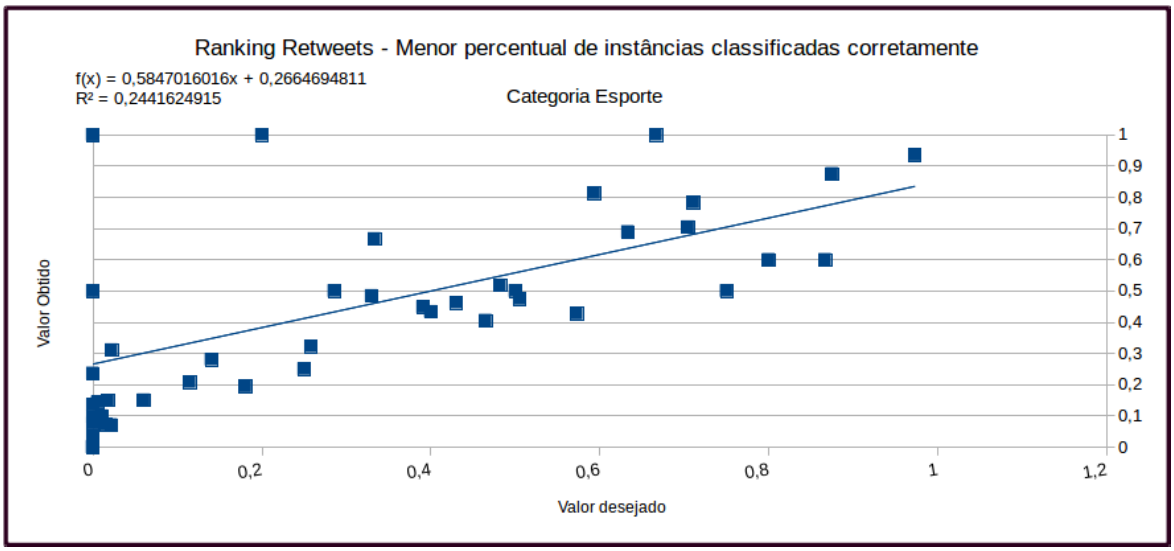
O percentual de pertencimento dos usuários a categoria *outros assuntos* é apresentado no Gráfico 21. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 1, 2, 3, 5, 6, 8, 10, 11, 12, 15, 20, 24, 25, 30, 31, 33, 38, 41, 42, 45, 46, 47, 50, 51, 52, 54, 57 e 58. A maior diferença apresentada entre estes valores foi de 1.00000.

Gráfico 21: Percentual de pertencimento a categoria *outros assuntos* dos usuários do *ranking* Retweets, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

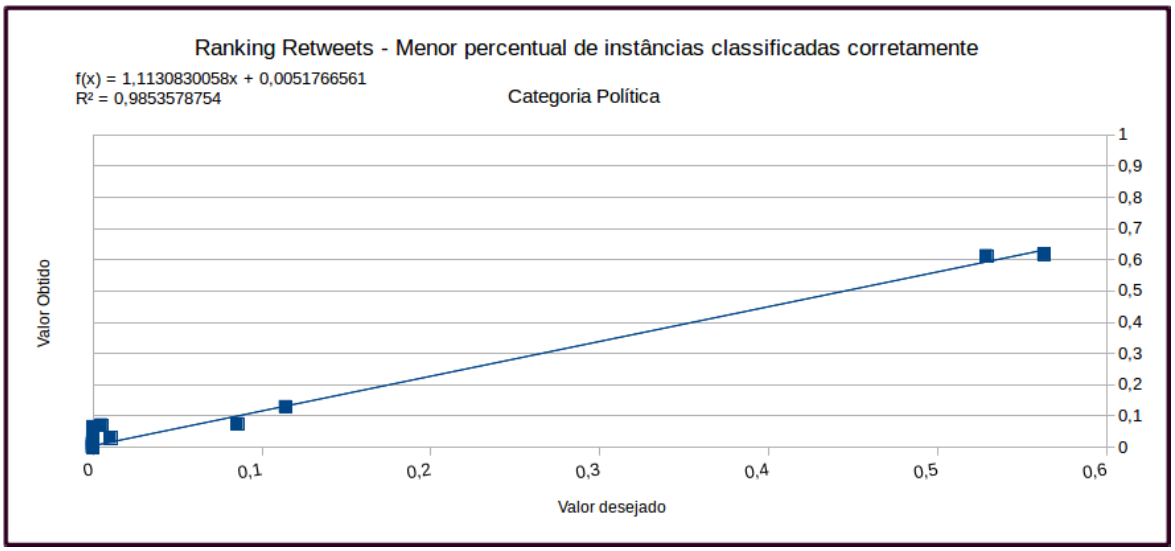
Gráfico 22: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *esporte*



Fonte: Produção do próprio autor.

No Gráfico 22 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *esporte*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.5847016016x + 0.2664694811$ e possui o coeficiente de determinação $R^2 = 0.2441624915$. Conclui-se então que 24.41624915% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

Gráfico 23: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *política*

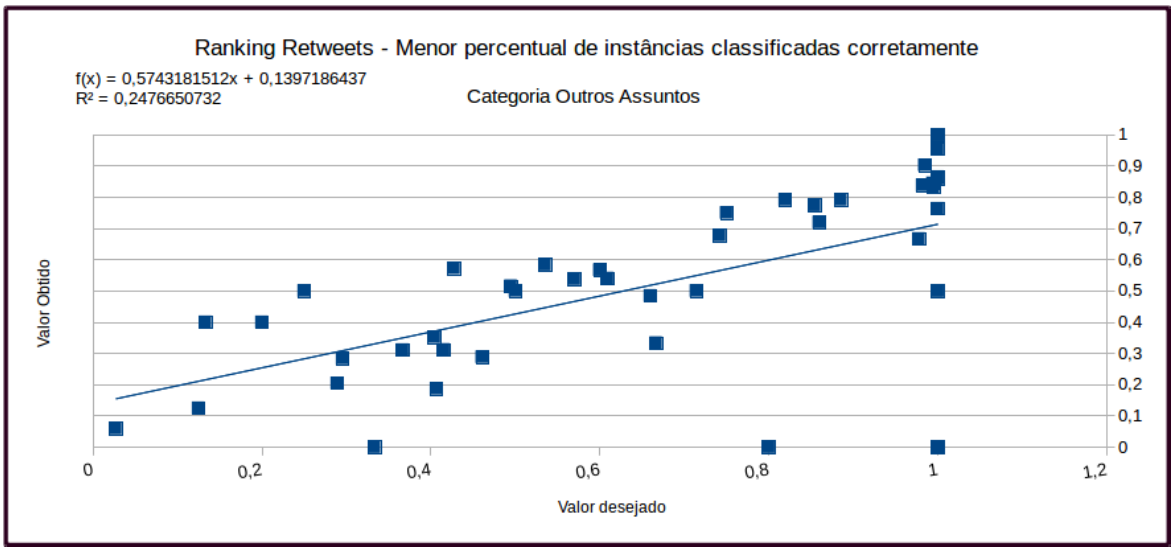


Fonte: Produção do próprio autor.

No Gráfico 23 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *política*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 1.1130830058x + 0.0051766561$ e possui o coeficiente de determinação $R^2 = 0.9853578754$. Assim temos que 98.53578754% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

No Gráfico 24 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *política*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.5743181512x + 0.1397186437$ e possui o coeficiente de determinação $R^2 = 0.2476650732$. Então, 24.76650732% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

Gráfico 24: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *outros assuntos*



Fonte: Produção do próprio autor.

A categoria *política* foi a que apresentou a menor diferença entre o percentual de postagens real e o obtido por usuário e a maior correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários. As categorias *esporte* e *outros assuntos* não ultrapassaram 25%.

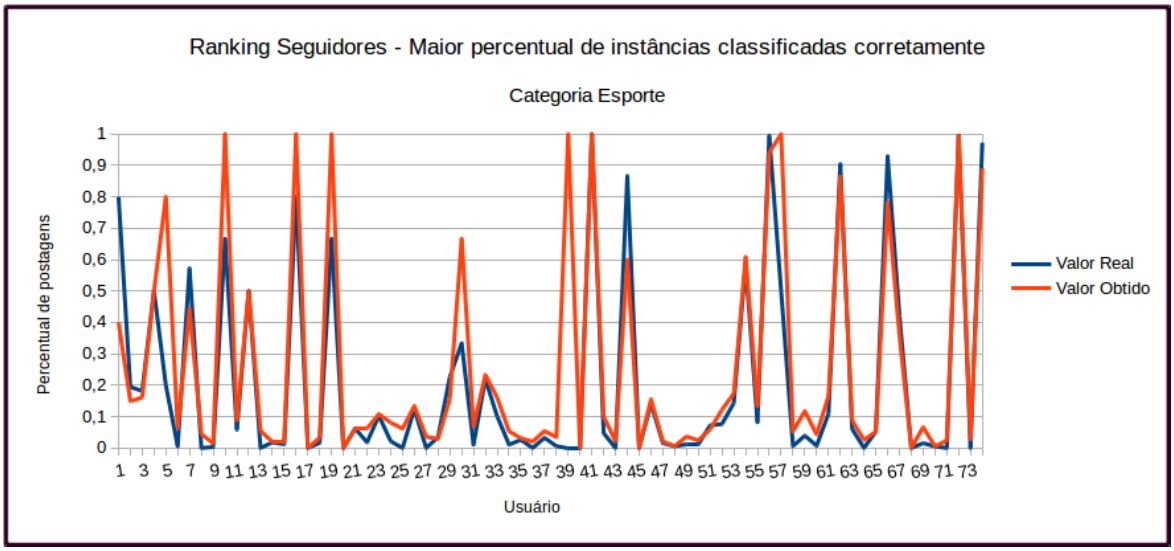
6.3.3 Ranking Seguidores

6.3.3.1 Maior percentual de instâncias classificadas corretamente

O quinto processo de categorização foi realizado com o resultado obtido do treinamento e classificação do *ranking* Seguidores que apresentou o maior percentual de instâncias classificadas corretamente.

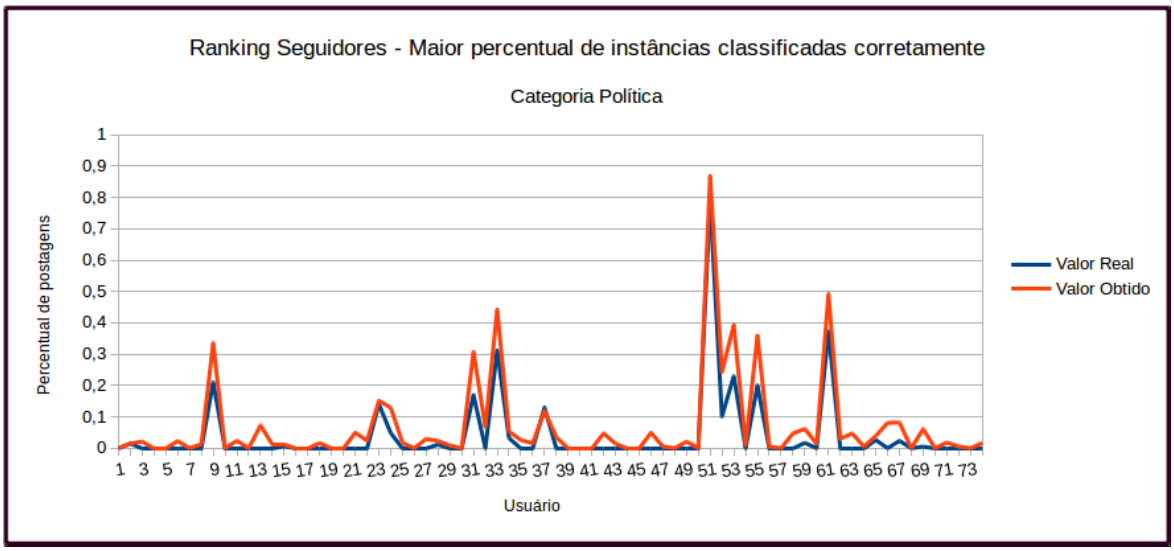
O percentual de pertencimento dos usuários a categoria *esporte* é apresentado no Gráfico 25. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 1, 5, 10, 16, 19, 22, 24, 25, 30, 39, 44, 57, 59, 66 e 74. A maior diferença apresentada entre estes valores foi de 1.00000.

Gráfico 25: Percentual de pertencimento a categoria *esporte* dos usuários do *ranking* Seguidores, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

Gráfico 26: Percentual de pertencimento a categoria *política* dos usuários do *ranking* Seguidores, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente

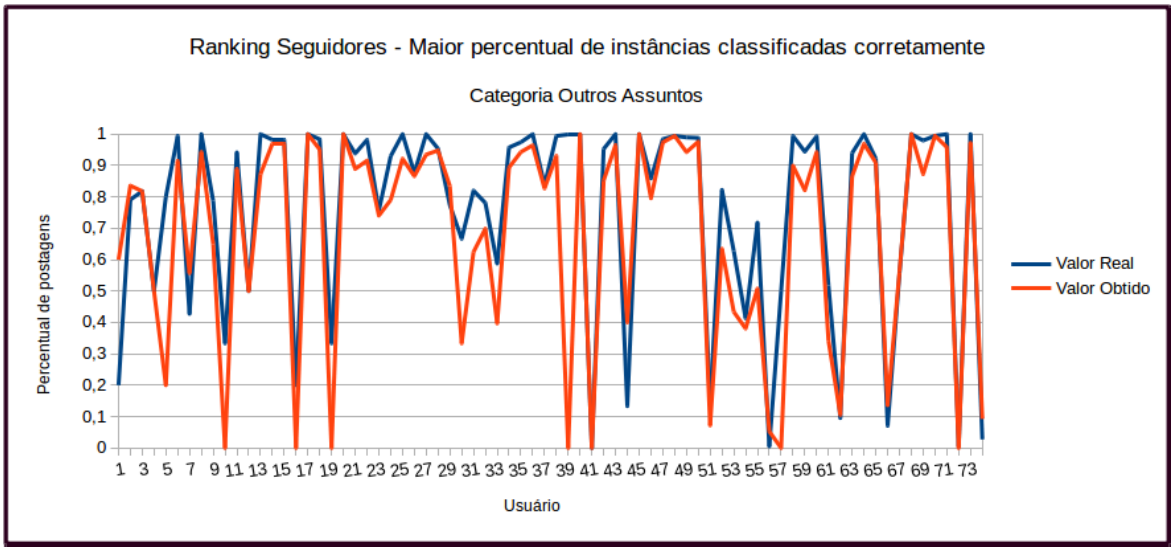


Fonte: Produção do próprio autor.

O percentual de pertencimento dos usuários a categoria *política* é apresentado no Gráfico 26. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 9, 13, 21, 24, 31, 33, 42, 46, 52, 53,

55, 58, 59, 62, 63, 66, 67 e 69. A maior diferença apresentada entre estes valores foi de 0.16327.

Gráfico 27: Percentual de pertencimento a categoria *outros assuntos* dos usuários do *ranking* Seguidores, baseando-se no treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente

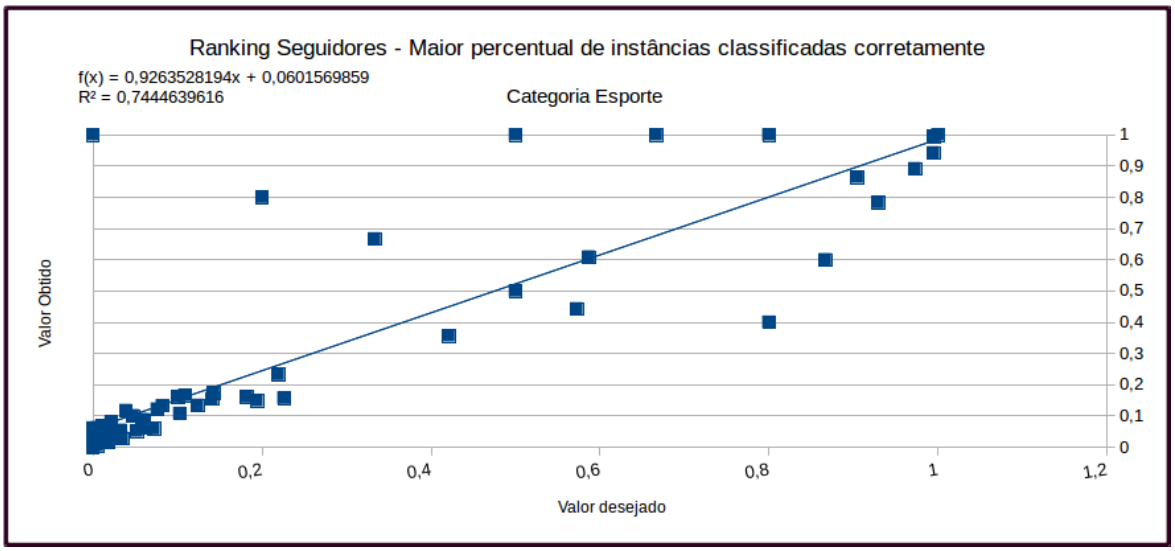


Fonte: Produção do próprio autor.

O percentual de pertencimento dos usuários a categoria *outros assuntos* é apresentado no Gráfico 27. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 1, 5, 6, 7, 10, 13, 16, 19, 21, 22, 24, 25, 27, 30, 31, 32, 33, 34, 39, 42, 44, 49, 52, 53, 55, 57, 58, 59, 69 e 74. A maior diferença apresentada entre estes valores foi de 1.00000.

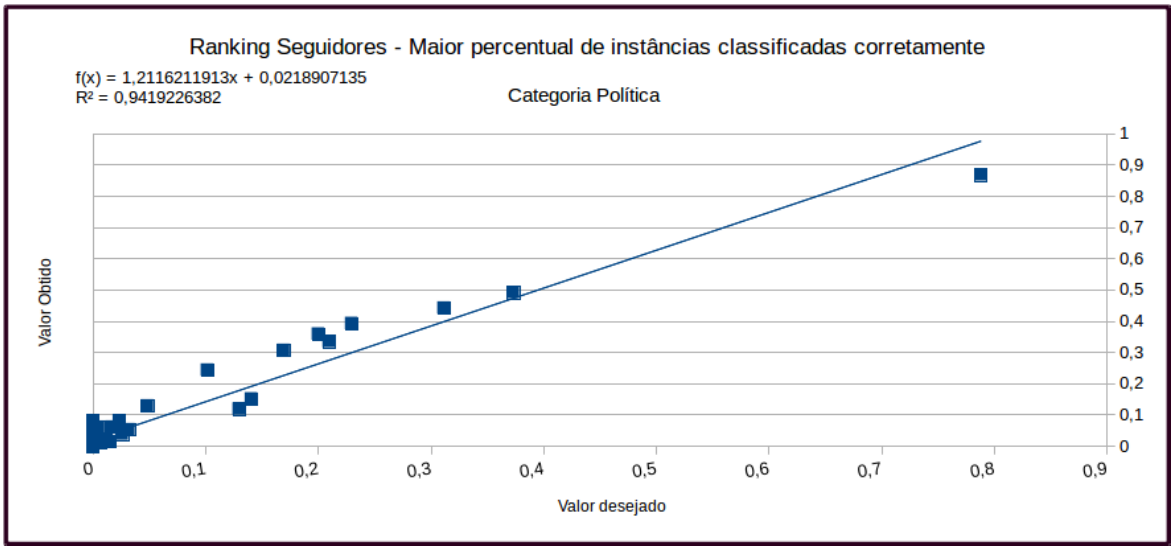
No Gráfico 28 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *esporte*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.9263528194x + 0.0601569859$ e possui o coeficiente de determinação $R^2 = 0.7444639616$. Conclui-se então que 74.44639616% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

Gráfico 28: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *esporte*



Fonte: Produção do próprio autor.

Gráfico 29: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *política*

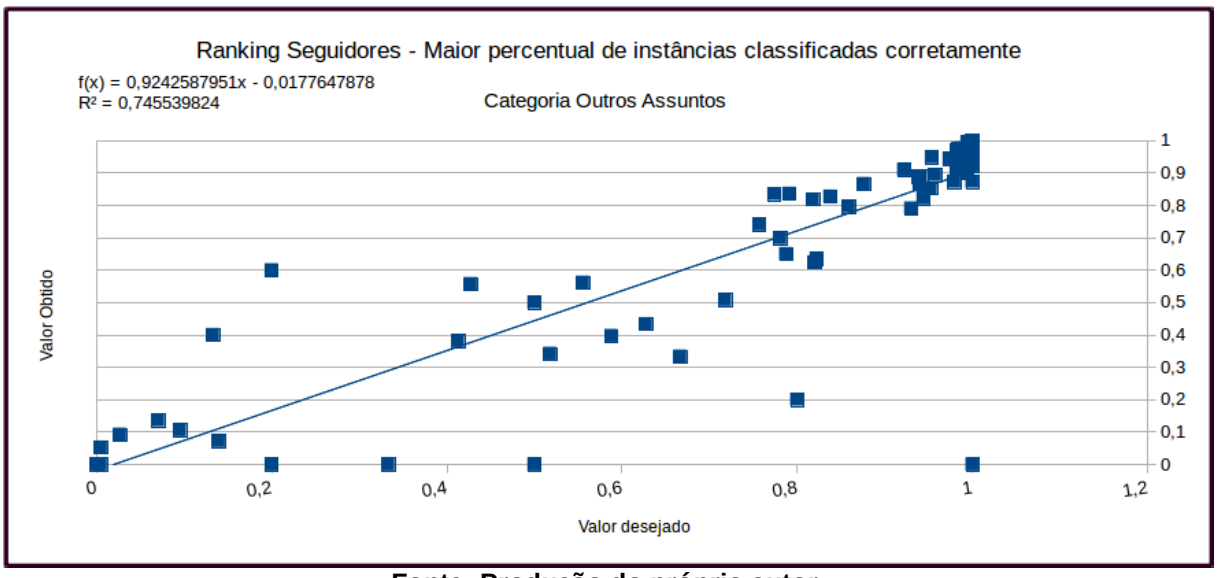


Fonte: Produção do próprio autor.

No Gráfico 29 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *política*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 1.2116211913x + 0.0218907135$ e possui o coeficiente de determinação $R^2 =$

0.9419226382. Logo, 94.19226382% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

Gráfico 30: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *outros assuntos*



Fonte: Produção do próprio autor.

No Gráfico 30 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *outros assuntos*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.9242587951x + 0.0177647878$ e possui o coeficiente de determinação $R^2 = 0.745539824$. Então, 74.5539824% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

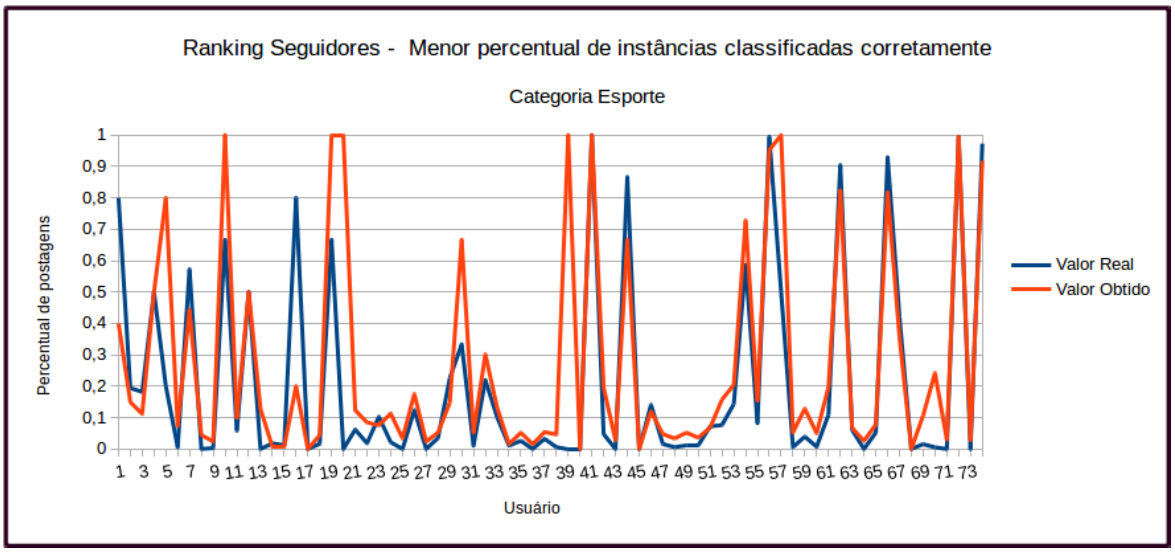
A categoria *política* foi a que apresentou a menor diferença entre o percentual de postagens real e o obtido por usuário e a maior correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários. As categorias *esporte* e *outros assuntos* obtiveram correlação superior a 74% entre os valores obtidos e os valores reais.

6.3.3.2 Menor percentual de instâncias classificadas corretamente

O último processo de categorização foi realizado com o resultado obtido do treinamento e classificação do *ranking* Seguidores que apresentou o menor percentual de instâncias classificadas corretamente.

O percentual de pertencimento dos usuários a categoria *esporte* é apresentado no Gráfico 31. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 1, 5, 6, 7, 10, 16, 19, 20, 21, 24, 30, 39, 42, 44, 54, 55, 57, 59, 62, 66, 69 e 70. A maior diferença apresentada entre estes valores foi de 1.00000.

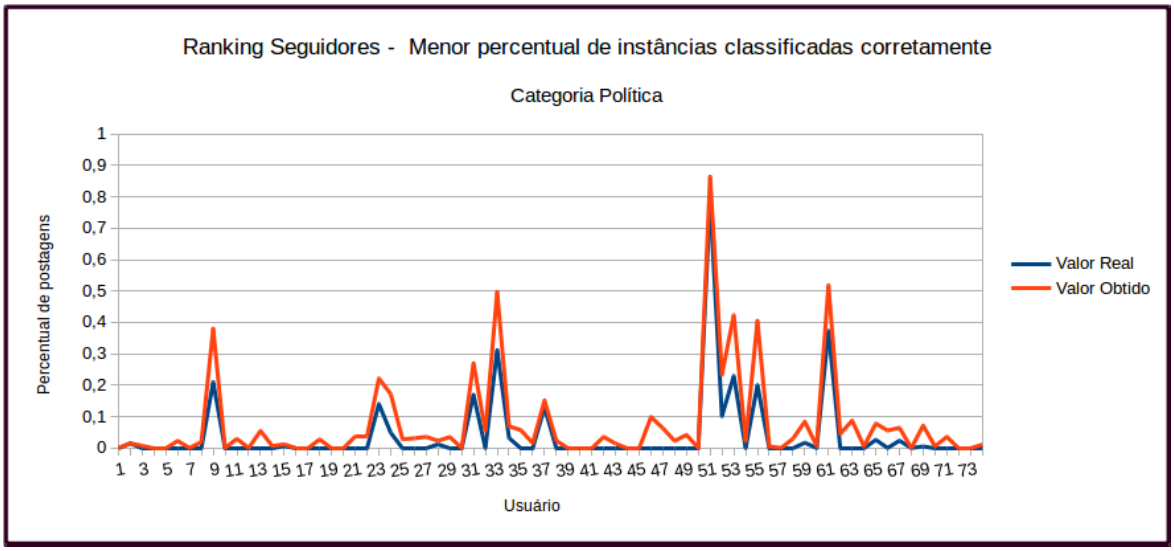
Gráfico 31: Percentual de pertencimento a categoria *esporte* dos usuários do *ranking* Seguidores, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

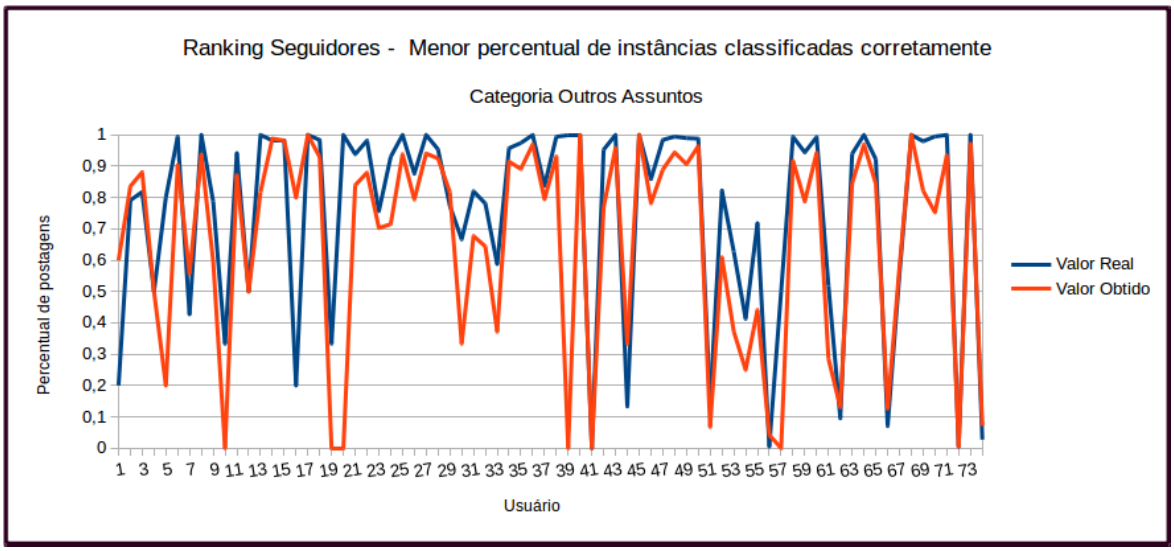
O percentual de pertencimento dos usuários a categoria *política* é apresentado no Gráfico 32. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 9, 13, 23, 24, 31, 33, 35, 46, 47, 52, 53, 55, 59, 63, 65, 66, 67, 69 e 71. A maior diferença apresentada entre estes valores foi de 0.20513.

Gráfico 32: Percentual de pertencimento a categoria *política* dos usuários do *ranking* Seguidores, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



Fonte: Produção do próprio autor.

Gráfico 33: Percentual de pertencimento a categoria *outros assuntos* dos usuários do *ranking* Seguidores, baseando-se no treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente



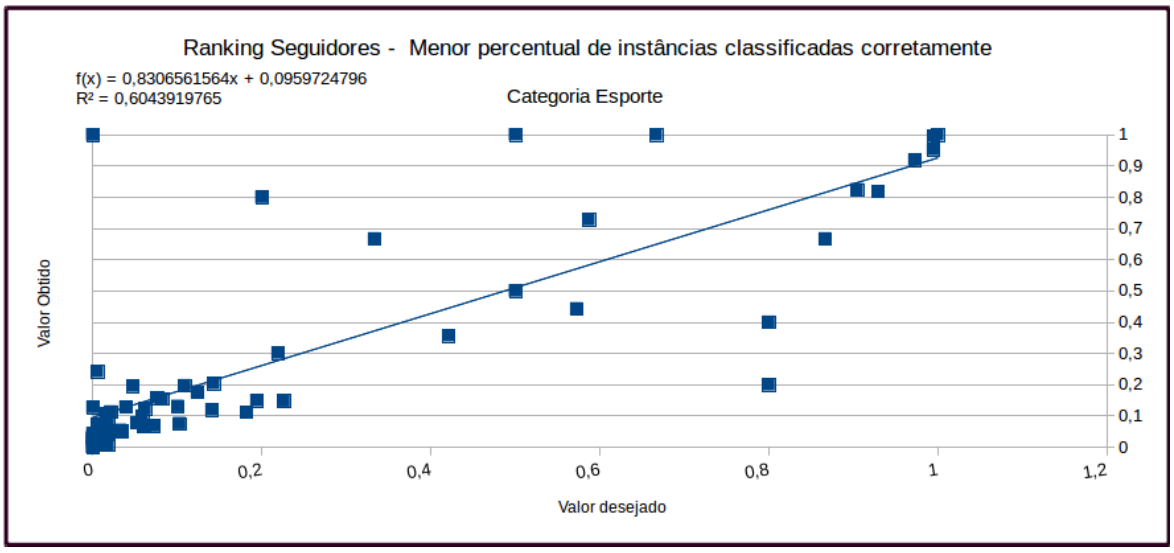
Fonte: Produção do próprio autor.

O percentual de pertencimento dos usuários a categoria *outros assuntos* é apresentado no Gráfico 33. As maiores diferenças entre os valores reais e os valores obtidos foram encontradas nos usuários: 1, 3, 5, 7, 10, 16, 19, 20, 21, 22, 23,

24, 25, 26, 27, 30, 31, 32, 33, 35, 39, 42, 44, 48, 52, 53, 54, 57, 59, 68, 69 e 70. A maior diferença apresentada entre estes valores foi de 1.00000.

No Gráfico 34 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *esporte*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.8306561564x + 0.0959724796$ e possui o coeficiente de determinação $R^2 = 0.6043919765$. Conclui-se então que 60.43919765% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

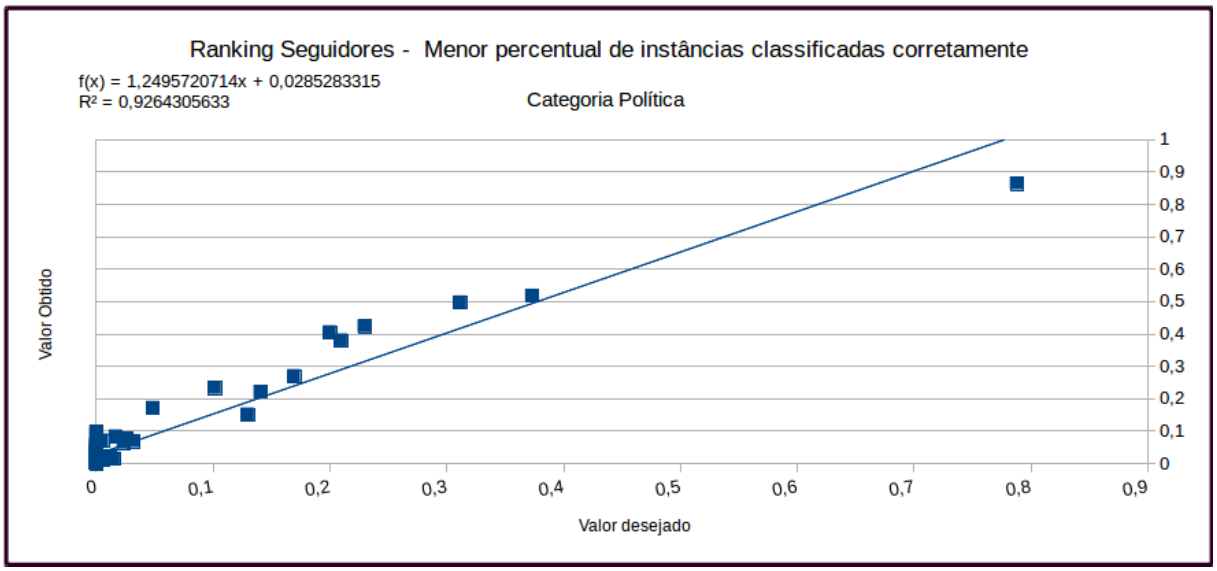
Gráfico 34: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *esporte*



Fonte: Produção do próprio autor.

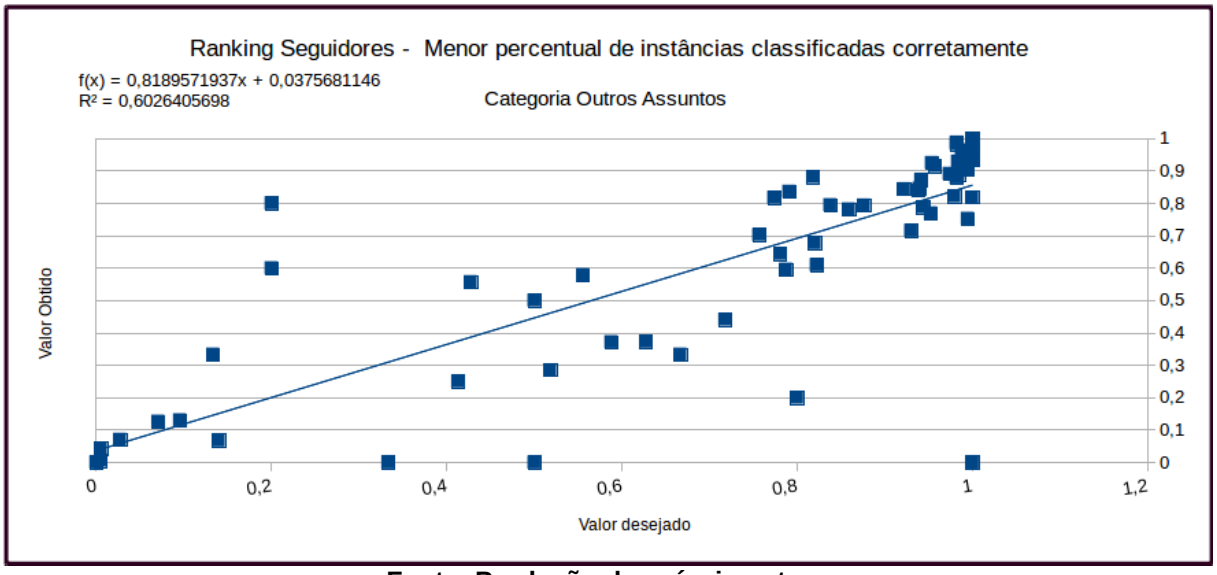
No Gráfico 35 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *política*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 1.2495720714x + 0.0285283315$ e possui o coeficiente de determinação $R^2 = 0.9264305633$. Então, 92.64305633% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

Gráfico 35: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *política*



Fonte: Produção do próprio autor.

Gráfico 36: Correlação entre o valor obtido e o valor desejado no percentual de pertencimento dos usuários à categoria *outros assuntos*



Fonte: Produção do próprio autor.

No Gráfico 36 têm-se a correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários à categoria *outros assuntos*. A reta de regressão que representa o modelo é expressa pela função $f(x) = 0.8189571937x + 0.0375681146$ e possui o coeficiente de determinação $R^2 =$

0.6026405698. Logo, 60.26405698% da variabilidade dos valores obtidos podem ser explicados pela variabilidade dos valores reais.

A categoria *política* foi a que apresentou a menor diferença entre o percentual de postagens real e o obtido por usuário e a maior correlação entre os valores obtidos e os valores reais no percentual de pertencimento dos usuários. As categorias *esporte* e *outros assuntos* obtiveram correlação superior a 60%.

7 CONCLUSÕES

Na etapa de classificação das postagens de cada *ranking* foi observado que os treinamentos e classificações que obtiveram um maior percentual de instâncias classificadas corretamente possuíam mais atributos do que os que obtiveram um menor percentual de instâncias classificadas.

Em todos os casos, os treinamentos e classificações que obtiveram um maior percentual de instâncias classificadas corretamente classificavam postagens da classe *outros assuntos*, enquanto os treinamentos e classificações que obtiveram um menor percentual de instâncias classificadas corretamente classificavam melhor postagens das classes *esporte* e *política*.

Na etapa de categorização dos usuários, pôde-se constatar que a categoria *política* foi a que apresentou o percentual de pertencimento do usuário mais próximo do valor real, em todos os processos realizados. Uma possível hipótese para isso consiste no conjunto de termos que representam essa categoria ser mais conciso do que nas demais categorias. O coeficiente de determinação médio que determina a correlação entre o percentual de pertencimento dos usuários obtido por esta categoria e o valor real foi de 0.9587088963 e o desvio padrão foi de 0.0232226162. O melhor coeficiente de determinação encontrado para esta categoria foi de 0.9853578754, valor conquistado na categorização do resultado do processo de treinamento e classificação que obteve o menor percentual de instâncias classificadas corretamente do *ranking Retweets*.

A categoria *esporte* obteve o coeficiente de determinação médio de 0.4632299917, com desvio padrão de 0.194783157. O maior coeficiente de determinação encontrado para esta categoria foi de 0.7444639616, valor conquistado na categorização do resultado do processo de treinamento e classificação que obteve o maior percentual de instâncias classificadas corretamente do *ranking* Seguidores.

A categoria *outros assuntos* obteve o coeficiente de determinação médio de 0.4708489375 e o desvio padrão de 0.1924865102 e o maior coeficiente de determinação encontrado para esta categoria foi de 0.745539824, valor conquistado na categorização do resultado do processo de treinamento e classificação que

obteve o maior percentual de instâncias classificadas corretamente do *ranking* Seguidores.

Num contexto geral, o processo de categorização realizado com o resultado do treinamento e classificação que obteve um maior percentual de instâncias classificadas corretamente no *ranking* Seguidores alcançou o melhor resultado.

Assim, conclui-se então que o processo de classificação das postagens realizado neste trabalho foi eficaz apenas para as postagens da categoria *política*.

Para continuidade deste trabalho sugere-se a modificação da etapa de classificação das postagens, criando um repositório prévio de palavras relacionadas a cada categoria, tendo assim um número fixo de atributos entre os treinamentos e classificações e possibilitando uma melhor comparação dos resultados com o de outras porcentagens de amostras selecionadas para treino.

Sugere-se também a implementação de uma técnica que limite a coleta dos usuários por localização geográfica, permitindo assim a identificação de usuários influentes por regiões definidas.

8 REFERÊNCIAS

BIGONHA, C. A. S. **Detecção de influência no Twitter baseada em sentimento**. 2012. 121f. Dissertação (Mestrado) - Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, 2012. Disponível em: < https://www.dcc.ufmg.br/pos/cursos/defesas/1484M.PDF >. Acesso em: 04 jul. 2015.

CARRILHO JÚNIOR, J. R. **Desenvolvimento de uma metodologia para mineração de textos**. 2007. 96 f. Dissertação (Mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: <http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=11675@1 >. Acesso em: 05 jul. 2015.

CHA *et al.* Measuring User Influence in Twitter: The Million Follower Fallacy. In: International AAAI Conference on Weblogs and Social Media, 4., 2010, Washington, DC, USA. **Proceedings…** Washington, DC, USA: AAAI, 2010. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1538%20Amit%20Goyal%2C%20Francesco%20Bonchi%2C%20Laks%20V.%20S.%20Lakshmanan%3A%20Approximation%20Analysis%20of%20Influence%20Spread%20in%20Social%20Networks%20CoRR%20abs/1826>. Acesso em: 04 jul. 2015.

CONECTA. 2014. **O jovem brasileiro na rede.** São Paulo, CONECTA, 36 transparências, color. Disponível em: <http://conecta-i.com/?q=pt-br/node/530>. Acesso em: 23 mar. 2015.

COSTA, J. A. F. Classificação automática e análise de dados por redes neurais auto-organizáveis. 1999. Tese (Engenharia elétrica) - Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, SP, 1999.

FACELLI, K.; LORENA, A.C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. Inteligência Artificial: Uma abordagem de Aprendizagem de Máquina. LTC: Rio de Janeiro, 2011.

FERREIRA, E. de B. A. **Análise de sentimento em redes sociais utilizando influência das palavras**. 2010. 69 f. Monografia (Graduação) - Universidade Federal de Pernambuco, Recife, 2010. Disponível em: < http://www.cin.ufpe.br/~tg/2010-2/ebaf.pdf >. Acesso em: 05 jul. 2015.

GARDELHA, R. N. de S. **Predição de Influência em Redes Sociais usando Traços de Personalidade**. 2013. 90 f. Dissertação (Mestrado) - Universidade Federal de Pernambuco, Recife, 2013. Disponível em: < http://repositorio.ufpe.br/bitstream/handle/123456789/12359/DISSERTA%C3%87AO%20Rene%20Gadelha.pdf?sequence=1&isAllowed=y >. Acesso em: 04 jul. 2015.

GONÇALVES et al. Métodos para Análise de Sentimentos no Twitter. In: Brazilian Symposium on Multimedia and the Web, 19., 2013, Salvador, Bahia, Brasil. **Proceedings…** Salvador, Bahia, Brasil: SBC, 2013. Disponível em:

<http://homepages.dcc.ufmg.br/~fabricio/download/webmedia13.pdf>. Acesso em: 19 maio 2015.

IBOPE. 2010. **Redes Sociais.** São Paulo, IBOPE Inteligência, v.1, 12 p. Disponível em: <http://www.ibope.com/pt-br/noticias/Documents/1008_WIN_redes_sociais.pdf>. (Relatório Técnico). Acesso em: 23 mar. 2015.

LIU, B. **Sentiment Analysis and Opinion Mining**. San Rafael, California: Morgan & Claypool Publishers, 2012. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.244.9480&rep=rep1&type=pdf>. Acesso em: 05 jul. 2015.

MITCHEL, T. M. Machine learning. McGraw-Hill: [S.l.], 1997.

MONARD, M. C. Aprendizado de máquina. In: MONARD, M. C. Sistemas inteligentes> fundamentos e aplicações. Manole: Barueri, São Paulo, Brasil, 2003.

MURPHY, K. P. **Machine Learning**: a probabilistic perspective. Cambridge: MIT Press, 2012. Disponível em: <http://www.cs.ubc.ca/~murphyk/MLbook/pml-intro-22may12.pdf>. Acesso em: 05 jul. 2015.

NASCIMENTO, P. et al. Análise de sentimento de tweets com foco em notícias. In Congresso da Sociedade Brasileira de Computação, 32., 2012, Curitiba, Brasil. Anais... Curitiba, Paraná, Brasil: SBC, 2012.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The pagerank citation ranking: Bringing order to the web. In: International World Wide Web Conference, 7., 1998, Brisbane, Australia. **Proceedings…** Brisbane, Australia: [s.n.], 1998. p. 161–172.

PARDO, T. A. S.; NUNES, Maria das Graças V. 2002. Aprendizado Bayesiano Aplicado ao Processamento de Línguas Naturais. São Paulo: USP, v.1, 26 p. Disponível em: <http://www.icmc.usp.br/~taspardo/NILCTR0225-PardoNunes.pdf>. (Relatório Técnico). Acesso em: 05 jul. 2015.

PINTO, C. M. S. **Algoritmos Incrementais para Aprendizagem Bayesiana**. 2005. 111 f. Tese (Doutorado) - Faculdade de Economia, Universidade do Porto, Porto, 2005. Disponível em: < http://w3.ualg.pt/~cpinto/tese.pdf >. Acesso em: 05 jul. 2015.

ROCHA, Eudson; ALVES, Lara Moreira. Publicidade online: o poder das mídias e redes sociais. **Fragmentos de Cultura**, Goiânia, v. 20, n. 3/4, p. 221-230, 2010.

SANTOS, L. M. et al. Twitter, análise de sentimento e desenvolvimento de produtos: quanto os usuários estão expressando suas opiniões?. Prisma.com, Porto, Portugal, v.1, n.13, p.1-12, 2010.

SANTOS, A. de M. Investigando a combinação de técnicas de aprendizado semissupervisionado e classificação hierárquica multirrótulo. 2012. 214 p. Tese (Ciência da computação) - Departamento de Informática e

Matemática Aplicada, Universidade Federal do Rio Grande do Norte, Natal, 2012.

SCHIMITT, V. F. **Uma análise comparativa de técnicas de aprendizagem de máquina para prever a popularidade de postagens no Facebook**. 2013. 57 f. Monografia (Graduação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: < http://www.lume.ufrgs.br/bitstream/handle/10183/86407/000910095.pdf?sequence=1 >. Acesso em: 05 jul. 2015.

SILVA, Gustavo Monteiro da. **Estudo e discussão da literatura sobre as principais técnicas para detecção de influência de usuários no twitter:** período 2010 a 2013. 2013. 56 f. Monografia (Graduação) - Universidade Federal de Lavras, Lavras, 2013. Disponível em: < http://repositorio.ufla.br/jspui/handle/1/5477 >. Acesso em: 18 maio 2015.

VALIATI, H., et al. Detecção de Conteúdo Relevante e Usuários Influentes no Twitter. In: Brazilian Workshop on Social Network Analysis and Mining, 1., 2012, Curitiba, Paraná, Brasil. **Proceedings...** Curitiba, Paraná, Brasil: SBC, 2012. Disponível em: <http://homepages.dcc.ufmg.br/~herico/brasnam.pdf >. Acesso em: 19 maio 2015.

WENG, J. et al. Twitterrank: Finding Topic-Sensitive Influential Twitterers. In: ACM International Conference on Web Search and Data Mining, 3., 2010. **Proceedings...** New York, ACM, 2010, p.261-270.