Revista de Sistemas de Informação da FSMA
n. 21 (2018) pp. 11-19

SISTEMAS DE INFORMAÇÃO
http://www.fsma.edu.br/si/sistemas.html

TRILHA PRINCIPAL ERI·GO

Detecção de *bullying*: Como identificar automaticamente essa prática em redes sociais?

Gaspar Monteiro e Silva*, Nádia F. F. da Silva*, Márcio de Souza Dias†
*Instituto de Informática - Universidade Federal de Goiás
† Departamento de Computação - Universidade Federal de Goiás - Regional Catalão
gaspar.monteiro@gmail.com, nadia@inf.ufg.br, marciosouzadias@ufg.br

***Resumo*—Técnicas de aprendizado de máquina podem ser utilizadas para inferir automaticamente informações não triviais a partir de grandes quantidades de dados. As redes sociais vêm ganhando popularidade e tornando-se importantes fontes de dados para a aplicação de tais técnicas computacionais. O objetivo deste trabalho é analisar técnicas de pré-processamento, conjuntos de atributos e classificadores aplicados à tarefa de detecção de traços de *bullying*, assim como a identificaçao do papel do autor do texto em relação ao episódio relatado. O trabalho foca em textos em Português do Brasil retirados de redes sociais. Vários classificadores e conjuntos de atributos foram estudados e comparados, com o objetivo de identificar qual é o mais apropriado para esta tarefa. De todas as configurações testadas, os melhores resultados obtidos para ambas as tarefas foram encontradas com o uso do maior conjunto de treinamento transformado em conjuntos de atributos compostos por unigramas e bigramas em conjunção com SVM utilizando kernel RBF.**

***Palavras-chave*—Bullying, Aprendizado de Máquina, Aprendizado Supervisionado, Twitter.**

Bullying Detection: How to automatically identify this practice in social networks?

***Abstract*—Machine learning techniques can be used to automatically infer information that is not available from large volumes of data. Social networks have been gaining popularity and turning into an important sources of data for an application of computer techniques. The goal of this work is to study how well preprocessing techniques, feature sets and classifiers work on the task of automatic bullying trace detection, as well as the role of the author of the text on the reported episode. We focused on social networks texts written in Brazilian Portuguese. Several different classifiers and attribute sets were studied and compared, in order to identify which one is the most appropriate for this task. Among all tested configurations, the best results was found when using the largest trainning set transformed into a feature set made of unigrams and bigrams in conjunction with SVM with an RBF kernel**

***Index Terms*—Bullying, Machine Learning, Supervised Learning, Twitter.**

I. Introdução

A prática de *bullying* consiste em ações negativas persistentes por parte de colegas (daí seu outro nome, *peer victimization*), sejam eles de escola, de trabalho ou de qualquer outro ambiente social [1]. Este assédio pode ser físico, verbal ou relacional [2], [3]. Os episódios geralmente são face a face, porém instâncias de *bullying* em meios virtuais vêm gradativamente se proliferando, estimuladas principalmente pelas interações entre usuários de redes sociais, sendo tais episódios conhecidos como *cyberbullying* [4].

As consequências do *bullying* ou *cyberbullying* para as vítimas vão de um maior risco de desenvolver doenças relacionadas ao estresse, males psicológicos, como transtornos emocionais e de personalidade, a até mesmo, em casos extremos, suicídio [5].

Para a sociedade como um todo, o *bullying* pode implicar em episódios de violência, como tiroteios em escolas [6] que são particularmente comuns nos Estados Unidos. No Brasil também foram notificados casos de tiroteios em escolas motivados de alguma forma por *bullying*, como o Massacre de Realengo [7] e o caso do atirador de Taiúva [8].

Em 2009, uma pesquisa realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), denominada de Pesquisa Nacional de Saúde Escolar [9], concluiu que 5.4% dos estudantes relataram ter sofrido *bullying* com alta frequência nos últimos 30 dias, enquanto que 25.4% mencionaram episódios com uma periodicidade menor.

A lei N^o13.185 de 6 de novembro de 2015 [10] foi instituída no Brasil para tentar combater o *bullying* em escolas e agremiações, e um dos seus objetivos é integrar os meios de comunicação de massa com as escolas e a sociedade, como forma de identificação e conscientização do problema. As redes sociais, um desses meios de comunicação de massa, são veículos para persuasão e formação de opiniões, sendo fortemente usadas para difusão de conteúdo.

Utilizar da grande quantidade de dados disponíveis nas redes sociais para estudar casos de *bullying* é um campo de estudo promissor, apesar de ainda pouco explorado. O crescimento da *Internet* e do conteúdo que é gerado por seus usuários, que publicam suas opiniões em uma linguagem coloquial e em muitos casos utilizando de artifícios

gráficos para tornar ainda mais sucintos seus diálogos é um importante desafio. Esse cenário é observado no Twitter[1], uma ferramenta de comunicação que pode facilmente ser usada como fonte de informação para ferramentas automáticas de detecção de *bullying*[11][12].

Os trabalhos encontrados na literatura ([13] e [14]) abordam o problema para língua inglesa sob o ponto de vista de algoritmos de Aprendizado de Máquina Supervisionados (classificadores), e não apresentam comparações ou extensões para outros idiomas, como por exemplo o Português do Brasil.

Outros trabalhos atacam o problema para a língua Portuguesa, porém lidam somente com *Cyber Bullying*[15] ou com outros problemas, como discurso de ódio [16].

Desafios importantes de pesquisa tanto do ponto de vista de Processamento de Língua Natural(PLN), quanto da aplicação de técnicas de Aprendizagem de Máquina(AM) e Análise Léxica(AL) são inerentes à língua. Por este motivo, este artigo visa nortear futuros trabalhos para a língua portuguesa que utilizem o Twitter, ou mesmo outras redes sociais para o estudo de *bullying*.

Especificamente, é possível enumerar como contribuições deste estudo: (i) a criação de um córpus, construído a partir do Twitter, contendo textos que possuam traços de *bullying* para o português, bem como a identificação do papel do escritor neste texto (podendo o autor ter o papel de vítima, relator ou praticante do ato de *bullying*); (ii) um estudo comparativo de diferentes classificadores para identificar traços de *bullying* em redes sociais; e (iii) uma vez identificado que tal texto se trata de um *traço de bullying*, prover a identificação automática do papel do escritor (classificar o autor do texto em: vítima, praticante ou relator/narrador de um episódio de *bullying*).

Este trabalho está organizado conforme segue: Na Seção II são definidos os papéis das pessoas envolvidas em episódios de *bullying*, bem como a fundamentação teórica. Na Seção III são apresentados os principais trabalhos relacionados. Na Seção IV, é apresentado brevemente como o córpus foi criado, na Seção V os experimentos realizados, os métodos e ferramentas utilizadas. Os resultados dos experimentos são relatados na Seção VI. A Seção VII apresenta algumas considerações finais e indicações de trabalhos futuros.

II. Identificação de traços de *bullying*

Neste trabalho, entende-se como traço de *bullying* um texto que mencione um episódio de *bullying* sofrido por alguém, ou um indivíduo relatando que alguém sofreu *bullying* em algum momento (mesmo que não detalhe o episódio), ou ainda instâncias de *cyberbulling*.

Esta definição é necessária pois é comum usuários de redes sociais se referirem a *bullying* de forma jocosa ou irônica.

Há também casos onde o termo *bullying* é usado de forma diferente das definições por este trabalho adotadas. Por exemplo, o *tweet "Agora q o Mexico tá sofrendo bullying econômico do Trump, ele vai procurar um psicólogo e depois desabafar fazendo textao no facebook"*, menciona que o atual presidente americano estaria fazendo *bullying* com o México. Definimos acima *bullying* como uma ação de indivíduos sobre indivíduos, e portanto este *tweet* não contém traço de *bullying*.

[1]https://twitter.com/

A. Papéis presentes em um episódio de bullying

Os indivíduos que participam de um episódio de *bullying* possuem papéis específicos [13]. A classificação de participantes segundo estes papéis é importante, pois cada participante do grupo tem uma função distinta no episódio, e a junção de todos os envolvidos é o que torna a prática de *bullying* possível. Assim, com o conhecimento destes papéis é possível trabalhar e contribuir para o combate ao *bullying* efetivamente [17].

Salmivalli, em [17], listou quatro papéis dos participantes de um episódio de *bullying* em adição ao ***Bully*** (plural ***Bullies***) e a ***Vítima***: (i) ***Assistente***, (ii) ***Reforçador***, (iii) ***Espectador*** e (iv) ***Defensor***. Para o autor, ***bullies*** são os indivíduos que ativamente praticam a ação. Enquanto que as ***vítimas*** são os indivíduos aos quais a ação é direcionada. Os ***assistentes*** são indivíduos que não iniciam um episódio, mas se juntam ao *bully* no ato. ***Reforçadores*** não se juntam diretamente ao episódio, mas reforçam os *bullies* através de *feedback* positivo (Rindo, por exemplo). Os ***espectadores*** presenciam o episódio mas não participam do mesmo. Enquanto que os ***defensores*** ativamente ajudam as vítimas, confrontando *bullies* e confortando a *vítima*.

Em [13], os autores adicionaram dois papéis quanto ao estudo de *bullying* em redes sociais: ***Relator*** e ***Acusador***. Segundo os autores estes papéis são necessários pois em muitos casos não fica claro se o escritor do *tweet* é uma vítima, espectador ou defensor. Os **Relatores** relatam um episódio ocorrido, podendo estar presentes ou não durante o episódio, enquanto que os **Acusadores** acusam alguém de cometer *bullying*.

Dada a complexidade de identificação dos papéis listados em [13],[17], principalmente pela natureza subjetiva da tarefa de anotação neste cenário (além da dificuldade de se identificar *tweets* com papéis muito específicos), neste trabalho são considerados apenas os papéis mais comuns: **Vítima**, **Bully** e **Relator** (veja Figura 1).

III. Trabalhos relacionados

Como mencionado anteriormente, [13] e [14] abordam o problema de detecção de traços de *bullying* para a língua inglesa. Os autores anotaram um córpus contendo *tweets* que potencialmente possuíam traços de *bullying* e usaram ferramentas de Processamento de Linguagem Natural e de Aprendizado de Máquina para obter um classificador para detectar automaticamente instâncias de *bullying*. Usando SVM linear juntamente

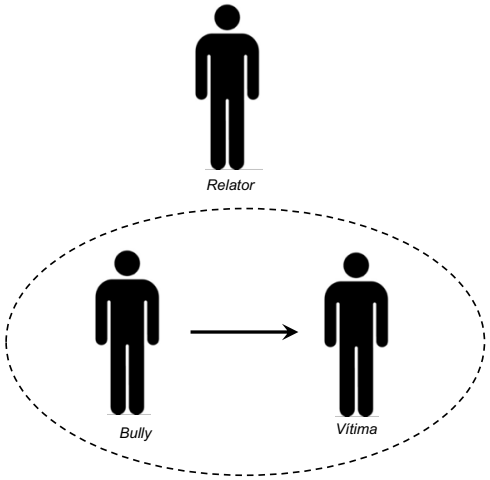

Fig. 1. Um episódio de *bullying*

com *unigramas e bigramas*[2], foi obtido uma acurácia de 81.3% na detecção de traços de *bullying* para a língua inglesa. É importante citar que não foi feito nenhuma radicalização de palavras(utilização das partes básicas de uma palavra) e as *stopwords*(palavras sem conteúdo, por exemplo, preposição, conjunção etc) não foram removidas. O córpus utilizado continha 1.762 *tweets*, sendo destes, 39% anotados com a classe "contém traços de *bullying*" e 61% com a classe "não contém traços de *bullying*". Outro experimento realizado pelos autores destes artigos foi classificar os *tweets* conforme o papel do escritor do *tweet*. Usando SVM linear juntamente com unigramas e bigramas, o classificador obtido alcançou uma acurácia de 61%.

Em [18], o autor realizou a coleta de *tweets* relacionados a professores, usando os termos de pesquisa "meu professor" e "minha professora". Após pré-processar os textos, 300 *tweets* foram anotados como Positivo, Negativo ou Neutro em relação aos professores. Estes *tweets* anotados foram usados para treinar um classificador usando Naive Bayes [19]. Tal classificador obteve uma taxa de acurácia de 87.1%. Apesar do artigo não explicitar como foi feita a anotação e não prover uma matriz de confusão, além de ter um conjunto de dados relativamente pequeno (300 *tweets*), os resultados do classificador se mostraram promissores o bastante para demonstrar que usar classificadores ao lidar com *tweets* em português pode ser uma alternativa viável para a análise de *bullying*.

Em [20], os autores utilizam de um grande volume de *tweets* sobre *bullying* para tentar responder algumas questões: Quem são os autores de relatos? O que eles relatam? Porque o fazem? De onde partem os relatos e quando eles são feitos? Foram analisados 9.764.583 *tweets* usando técnicas de aprendizado de máquina para prover respostas a estas perguntas. Usando LDA juntamente com SVM, os autores obteram precisão de 89% recall de 85.5% e f-score de 87%.

As redes sociais criaram um ambiente propício ao *cyberbullying*. Logo vários estudos foram feitos para entender melhor como o *cyberbullying* funciona. Em [21], os autores utilizam técnicas de aprendizado de máquina para detectar automaticamente instâncias de *cyberbullying* no Twitter utilizando um esquema de peso baseado em seleção de atributos (*feature selection*) juntamente com latent Dirichlet allocation (LDA).

Em [22], os autores utilizam além dos textos dos *tweets*, metadados, como número de seguidores, número médio de postagens diárias, data da última vez postada, data da criação de conta, dentre outros, para extrair características de *cyberbullies* no Twitter. Entre as conclusões apresentadas, estão o fato de que usuários agressivos mostram um comportamento similar ao de *Spammers*(usuários que enviam grandes quantidades de mensagens repetidas, geralmente de conteúdo mercadológico), por exemplo, em termos de número de seguidores, amigos e tamanho da rede. Tais usuários não postam muitos *tweets* e usam poucas *hashtags*(#) e URLs em seus *tweets*. Além disso, estes usuários costumam estar no Twitter há pouco tempo. Por outro lado, diferente dos *Spammers*, os usuários comuns são bastante populares, considerando-se número de seguidores, amigos e tamanho da rede. Eles participam de muitas comunidades e usam um grande número de *hashtags* e URLs.

Os trabalhos relacionados apresentados nesta seção evidenciam a grande relevância do objeto de estudo neste artigo. Assim, para a língua portuguesa, um dos maiores desafios é prover um córpus que funcionará como um *benchmark* para detecção de traços de *bullying* em Português, bem como análises fundamentadas e comparativas dos principais algoritmos de classificação aplicados à estes dados, e cuja a acurácia já foi comprovada para a língua inglesa, mas não comprovada para o Português do Brasil.

IV. Etapas para Classificação de Traços de *bullying*

Os métodos de detecção de traços de *bullying* seguem uma sequência de passos comuns à área de mineração de textos [13], [14], conforme descrito na Figura 2. Vamos descrever como implementamos cada um destes passos na subseções a seguir.

Pessoas que presenciam ou vivenciam um episódio de *bullying* comumente reportam ou comentam o mesmo em redes sociais, o que faz da extração de traços de *bullying* em redes sociais possível.

O Twitter foi a rede social escolhida para este estudo, pois apesar de suas limitações em número de caracteres, a mesma possui várias vantagens. Sua API[3] permite facilmente coletar uma grande quantidade de dados usando palavras chave e além disso é possível monitorar trocas

[2]Unigramas são termos compostos por apenas uma palavra extraídos dos textos utilizados nas análises, enquanto que bigramas são termos que possuem duas palavras.

[3]https://dev.twitter.com/rest/public

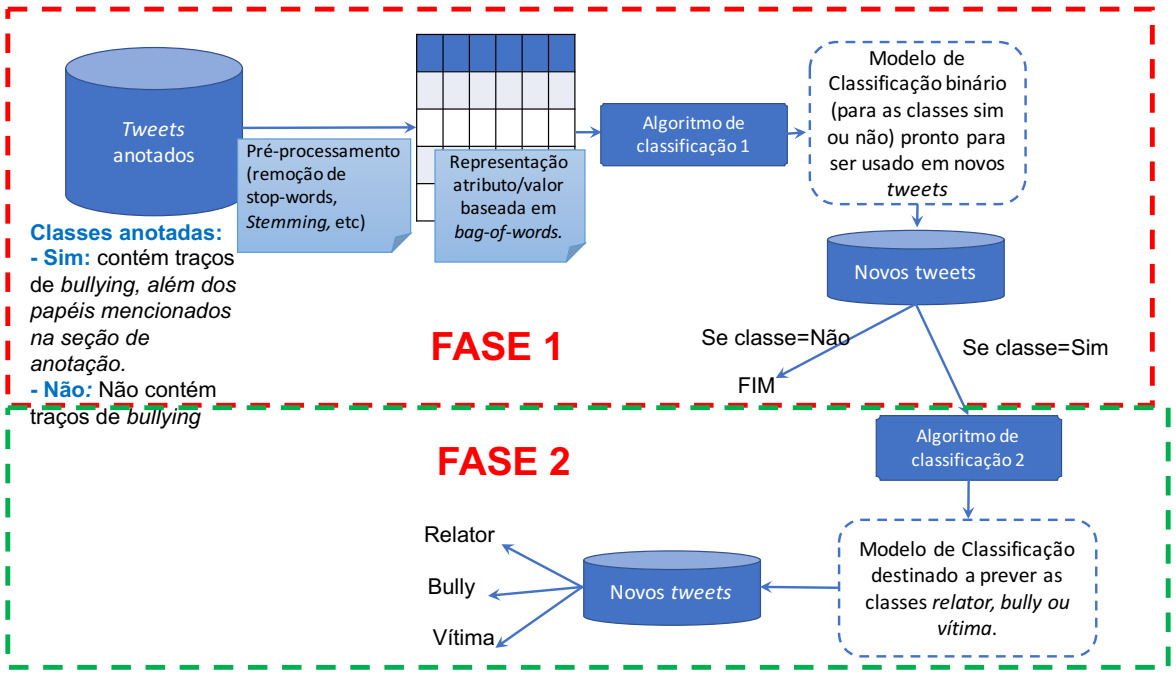

Fig. 2. Etapas realizadas para a obtenção de um modelo de classificação de detecção de traços de *bullying*. Na Fase 1 são feitos os pré-processamentos necessários e a conversão dos textos para representação vetorial, assim como a classificação do *tweet* quanto a presença ou não de traços de *bullying*. Os *tweet* identificados na classe "Sim" são então movidos para a Fase 2, onde são classificados quanto ao papel do autor.

de mensagem e viralizações através de *retweets*[4] e *replies*[5] – Caso um *tweet* seja resposta a outro, e este contenha traço de *bullying*, é possível analisar o *tweet* original. Pode-se também usar os dados da quantidade de *retweets* para fazer um estudo sobre viralização de *tweets* com traço de *bullying*, ou até mesmo o uso de *retweets* como reforçamento do *bullying*.

A. Construção do corpus

Para a construção do corpus, inicialmente foram coletados 48.262 *tweets* usando a API de *streaming* do Twitter. Os *tweets* foram coletados da seguinte forma: Todos os *tweets* que continham as palavras chave ***bullied, buly, bullying, bulin, buli, bullin, bulen, bullen e bulli*** foram recebidos por um *script* em *Python*, utilizando a biblioteca *tweepy*[6]. Foram adicionados as grafias incorretas dos termos em inglês pois os mesmos são frequentemente utilizados por usuários brasileiros. A coleta foi feita em tempo real, 24 horas por dia, entre 19 de Novembro de 2016 e 27 de Janeiro de 2017. Os *tweets* que não eram do Português do Brasil foram descartados. Nenhum *tweet* foi duplicado.

Após a coleta foram excluídos *tweets* que representavam *spams*. Foi percebido que as palavras chaves utilizadas traziam um grande numero de *tweets* com o mesmo texto: "Morgan Dollar Uncirculated US Mint Gem PQ Silver Coin BU Unc MS++++++ HTTPLINK #silver #junksilver #coins #us #bulli…". Foram removidas todas as ocorrências deste texto na base de dados. De forma a maximizar a quantidade de informação presente nos textos anotados, tendo em vista que somente 2000 *tweets* seriam anotados, também foram excluídos do conjunto de interesse *tweets* que possuíam menos de 120 caracteres. Após todos estes filtros restaram 6.851 *tweets* de interesse. A sequencia de passos de processamento e a quantidade de textos restantes após cada fase é mostrada na tabela I

TABELA I
DIFERENTES PASSOS NO PROCESSAMENTO

Base de dados	**Quantidade de textos**
Todos os textos	48.262
Após remoção de spams	47.231
Após Remoção de textos com menos de 120 caracteres	6.851
Seleção final para anotação	2.000

Deste montante, 2.000 *tweets* foram anotados por 6 anotadores quanto à presença de traço de *bullying* e quanto ao papel do autor em relação ao episódio. Os 2.000 *tweets* foram divididos em dois grupos de 1.000, cada grupo sendo anotado por 3 anotadores.

De forma a facilitar o processo de anotação e de forma a contornar o problema da subjetividade da anotação (pode haver discordância quanto ao que é ou não um episódio de *bullying*), foi confecionado um manual para os anotadores, contendo as motivações para o trabalho, diretrizes para anotação, e principalmente uma definição objetiva do que deveria ser considerado como um traço de *bullying*. Os anotadores deveriam então, seguir como base na seguinte

[4]Os *retweets* ou RTs são mensagens repassadas de algum usuário no Twitter para os seus seguidores.

[5]*Replies* são mensagens direcionadas a algum usuário referenciando o nome do usuário.

[6]http://www.tweepy.org

definição: "Um *tweet* possui traço de *bullying* quando há menção a um ou mais episódios de *bullying*, que acuse alguém de ser um *bully* sem mencionar episódios específicos, que mencione que o autor ou outra pessoa sofreu *bullying*, mesmo que sem mencionar episódios específicos ou instâncias de *cyberbullying*."

Os anotadores receberam um arquivo contendo o número de identificação (ID) e o texto do *tweet*, um campo chamado *Bullying* (onde o anotador deveria anotar "sim" caso fosse identificado um traço de *bullying* e "não" caso contrário). Além disso, também foi disponibilizado no arquivo o campo Papel para que o anotador (quando anotar o campo *Bullying* com o valor "sim") identifique se o traço de *bullying* foi escrito por uma vítima, por um *bully* (praticante do episódio de *bullying*) ou por um relator (narrador do episódio de *bullying*). O arquivo final após a anotação é ilustrado na Tabela II. O arquivo com os IDs e o resultado das anotações foi colocado no Github[7].

V. Experimentos

Uma vez de posse do corpus anotado, o próximo passo deste estudo foi avaliar algoritmos de classificação clássicos já avaliados para outras línguas para o problema de detecção de traços de *bullying*. Neste trabalho foram usados os resultados da anotação do corpus integralmente, ou seja, os textos anotados relativos a ambos os grupos. O total de textos anotados foi de 2.000 *tweets*. Destes 2.000 textos, 1.294 foram rotulados como contendo traços de *bullying*, e 706 como não contendo traços de *bullying*. Quanto ao papel do autor, 102 foram rotulados como *Bully*, 442 como Vítima e 629 como Relator. Houve discordância total entre os 3 anotadores quanto ao papel em 121 exemplos, estes então não foram alocados em nenhuma classe.

Apesar de o *dataset* incluir apenas *tweets* que continham os termos relacionados a *bullying*, várias instâncias não contém traço de *bullying* efetivamente, como por exemplo: "*@USERNAME COMASSIM? SÓ PODE QUEM TIVER 18 EM 2020? Q BULLYING! eu vou ter 16 ;-; LETÍCIA SE TU FOR, ME LEVA MOZONA parei*".

O resultado esperado de um algoritmo de classificação supervisionado é um modelo que será usado para classificar novas instâncias (diferentes das instâncias usadas no treinamento) conforme classes predeterminadas. Este modelo, também conhecido como classificador, pode ser construído usando várias técnicas diferentes. Neste trabalho, foram usados os classificadores Naive Bayes, Regressão Logística (*Logistic Regression*) e SVM (*Support Vector Machine*)[19]. Ao usar SVM, é necessário escolher um *kernel*. Neste trabalho foram utilizados o *kernel* linear e o *kernel* RBF [23]. Ao se utilizar o *kernel* RBF em SVM , é necessário calibrar alguns parâmetros internos do mesmo, nomeadamente *C* e *gamma*. Para este trabalho foi utilizada a abordagem de calibragem de parâmetros demonstrada em [24]. Esta abordagem consiste em rodar várias vezes o algorítimo de classificação com valores diferentes até chegar ao melhor valor. Após realizar a abordagem citada, os valores obtidos foram C = 8, gamma = 0.01 para a detecção de traços de *bullying*, e C = 3.75, gamma = 0.005 para a classificação quanto ao papel do autor.

Uma forma de comparar classificadores é comparar suas medidas de acurácia. A acurácia é definida pelo número de instâncias corretamente classificadas dividido pelo total de instâncias. Além da acurácia, outras medidas são utilizadas neste trabalho. Nomeadamente Precisão, *Recall* e *F-score*[19]. Estas medidas são especialmente úteis quando não se tem uma base de dados perfeitamente balanceada, ou seja, quando a quantidade de instâncias de cada classe não é igual.

Uma forma de visualizar o desempenho de um classificador é através da sua Matriz de Confusão. Uma Matriz de Confusão é uma matriz que exibe em um eixo as classes preditas pelo classificador, e em outro eixo as classes reais dos dados. Através da Matriz de Confusão é possível observar a eficácia do classificador em relação a cada classe dos dados.

Para executar os experimentos, criar os conjuntos de atributos e avaliar os classificadores obtidos, foi utilizada a linguagem Python[8]. Python possui um módulo específico para ciência de dados, chamado *scikit-learn* [25], que contém toda a implementação dos classificadores utilizados, funções de transformação de texto em atributos, separação de dados para teste e treinamento, e funções de avaliação.

Seguindo [13], executamos 30 rodadas de cada experimento, sendo os resultados a média aritmética das execuções. As descrições dos experimentos executados nas Fases 1 e 2 estão em seguida.

A. Fase 1 - Detectar se existe um traço de bullying no tweet

A Fase 1 foi realizada no conjunto de 2.000 tweets, dentre os quais, 300 (15%) foram separados como conjunto de teste. Os 1.700 tweets restantes foram utilizados para a etapa de treinamento. Realizamos um treinamento incremental, a fim de avaliar o quanto a quantidade de instâncias no corpus de treinamento impacta no resultado. Desta forma, utilizamos inicialmente 100 tweets para treinamento, gerando um modelo de classificação a ser avaliado nos tweets de teste. Após cada avaliação, foram acrescentados novos 100 tweets no conjunto de treinamento e posterior avaliação. Tal processo se repetiu por 17 vezes, até atingir a totalidade do conjunto de treinamento (mesma metodologia utilizada em [13]). Os conjuntos de teste e treinamento foram gerados de forma aleatória em cada execução, garantindo-se que a distribuição entre as classes fosse mantida em todas as execuções (estratificação).

Os textos foram transformados em dois conjuntos distintos de atributos. No primeiro conjunto, os textos foram transformados em *bag-of-words* ou um conjunto de atributos (ver Fase 1 da Figura 2) contendo somente unigramas. Já no segundo conjunto, os textos foram transformados

[7]https://github.com/makotohadou/bullyingIDS/blob/master/finalFileWholeDataset.csv

[8]https://www.python.org/

TABELA II
ARQUIVO DE ANOTAÇÃO

ID	**Texto**	**Bullying?** (Sim ou Não)	**Papel** (vítima, *bully* ou relator)
30	Já fui alvo de "bullying", na escola diziam que eu parecia um menino pq eu usava roupas da sessão de meninose. Eu odiava isso!	S	vítima
36	#SouDoTempoQue meninos eram simplesmente meninos, meninas eram meninas, todo mundo se zoava e ngm chorava "bullying mimimi"	N	N/A
45	Tava reclamando da festa, mas na real foi muito boa, encontrei amigos antigos, descobri q os garotos q eu fazia bullying n me odeiam	S	bully
6679	Aluno sofre bullying por ser careca, e diretor o deixa raspar sua cabeça em sala	S	relator
6717	esse povo que vitimiza o snape pq ele sofreu bullying na escola mas apaga tudo o que ele fez com os alunos da gryff...	S	relator
6742	vi em uma entrevista da pabllo falando o bullying que ela sofre por causa da por dela eu fiquei ????? a voz dela é maravilhosa gente	S	relator
6818	uma pagina de youtubers no fb postou que o Christian sofreu bullying por ser magro demais. O que é de total verdade, quem acompanha ele (+)	S	relator

em uma *bag-of-words* contendo unigramas e bigramas. Nenhum pré-processamento adicional foi empregado no conjunto de dados. Quanto à Fase 1, a *bag-of-words* contendo unigramas, após a extração, apresentou 7.387 termos (atributos), já a *bag-of-words* contendo unigramas e bigramas apresentou 35.542 termos.

Sumarizando, foram executadas 30 rodadas de testes para cada classificador e cada conjunto de atributos. Sendo que cada par classificador+conjunto de atributos foi testado com 17 tamanhos diferentes. O modelo foi inicialmente treinado em 100 *tweets* e o processo de treinamento foi repetido com incrementos de 100 (até atingir a totalidade do conjunto de treinamento). Ao final foram realizados 4.080 experimentos. A hipótese inicial com essa configuração experimental é que ao adicionar novos dados ao conjunto de treinamento ocorreria um ganho de acurácia gradativo, o que de fato pode ser visualizado nos gráficos das Figuras 3 e 4.

B. Fase 2 - Classificar o papel dos envolvidos no episódio de bullying

Para a Fase 2, somente 1.173 tweets foram utilizados, pois foram descartadas as instâncias anotadas como "não contendo traços de *bullying*" e as instâncias onde os anotadores discordaram sobre o papel do autor. Destas 1.173 instâncias, 173 (aproximadamente 15%) foram separadas como conjunto de teste. Apenas 102 *tweets* foram rotulados como *Bully*, 442 como Vítima e 629 como Relator.

Inicialmente foram replicados os mesmos experimentos realizados na Fase 1 para esta tarefa, considerando os mesmos atributos (unigramas e bigramas) e os mesmos algoritmos. Entretanto, com os atributos mencionados anteriormente, os resultados foram insatisfatórios. Buscando suavizar o problema da quantidade insuficiente de amostras, acrescentamos outros atributos:

1) *N-grams:* Foram considerados unigramas, bigramas e trigramas, os quais representam sequências formadas, respectivamente, por uma, duas e três palavras.
2) *Estilo de Escrita:* atributos foram derivados a partir da presença de três ou mais caracteres repetidos em palavras, da sequência de três ou mais pontuações e do número de palavras com todas as letras em maiúsculo.

VI. RESULTADOS

A. Fase 1 - Detectar se existe um traço de bullying em um tweet

Na Figura 3 os resultados para a Fase 1 são apresentados, utilizando apenas unigramas como atributos e os classificadores Naive Bayes, Regressão Logística, SVM com Kernel Linear e SVM com Kernel RBF.

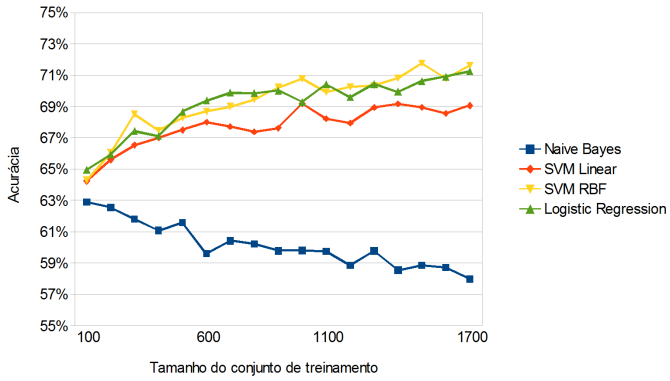

Fig. 3. Resultados dos experimentos da Fase 1 (Detectar se existe um traço de *bullying* no *tweet*) utilizando somente unigramas

Quanto à acurácia, o SVM com *kernel* RBF obteve os melhores resultados, chegando a 72 % no maior conjunto de dados. A Regressão Logística e o SVM com *kernel* Linear alcaçaram, 71.2% e 69% respectivamente. Naive Bayes foi significativamente pior, com 58% (ver Figura 3).

Na Figura 4, os resultados para a Fase 1 são apresentados, utilizando apenas unigramas e bigramas como atributos e os classificadores Naive Bayes, Regressão Logística, SVM com Kernel Linear e SVM com Kernel RBF.

Segundo a Figura 4, o SVM com *kernel* RBF obteve os melhores resultados, chegando a 72.8% no maior conjunto de dados. SVM com *kernel* Linear e Regressão Logística foram levemente piores, com 71.5% e 71.8% respectivamente. Naive Bayes obteve a acurácia mais baixa dentre os classificadores utilizados, 67.7%.

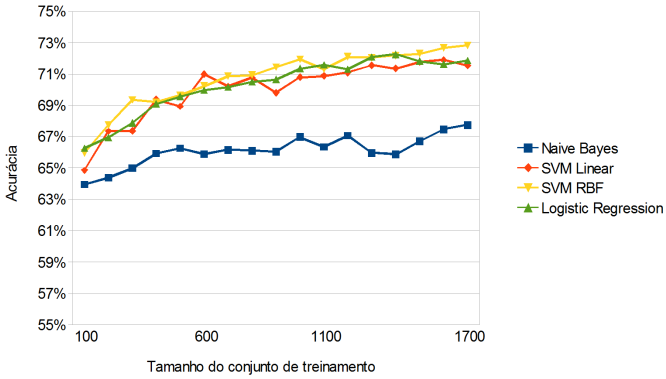

Fig. 4. Resultados dos experimentos da Fase 1 (Detectar se existe um traço de *bullying* no *tweet*) utilizando unigramas e bigramas

O classificador mais bem sucedido, SVM com *kernel* RBF, obteve a Precisão de 75%, *Recall* de 87% e F-score de 80%. Para ilustrar melhor os resultados, a matriz de confusão extraída deste classificador é mostrada na Tabela III.

Para esta tarefa, obtivemos um resultado abaixo do apresentado em [13], [14], onde os autores obtiveram resultados de 80% para esta tarefa na língua Inglesa. Isso pode ser devido à diferenças na língua usada, e também ao comportamento de usuários em relação a bullying em diferentes países.

TABELA III
MATRIZ DE CONFUSÃO UTILIZANDO SVM RBF COM ATRIBUTOS COMPOSTOS DE UNIGRAMAS E BIGRAMAS PARA OS DADOS DE TESTE - 300 *tweets* (FASE 1 - DETECTAR SE EXISTE UM TRAÇO DE *bullying* NO *tweet*).

		Classe real	
		Sim	Não
Classe prevista	Sim	163	54
	Não	31	52

Um ponto interessante a se notar é que utilizando apenas unigramas como atributos, o Naive Bayes teve seu desempenho deteriorado com o aumento do conjunto de treinamento. Acreditamos que tal comportamento tenha ocorrido em função do viés de aprendizado do algoritmo Naive Bayes, o qual considera as features independentes, e portanto captura pouca informação de contexto.

A Figura 5 mostra a diferença na acurácia entre os conjuntos de atributos, usando o classificador mais bem sucedido como exemplo.

Apesar dos resultados serem positivos, podemos enumerar possíveis motivos para a acurácia ainda estar longe dos 100%:

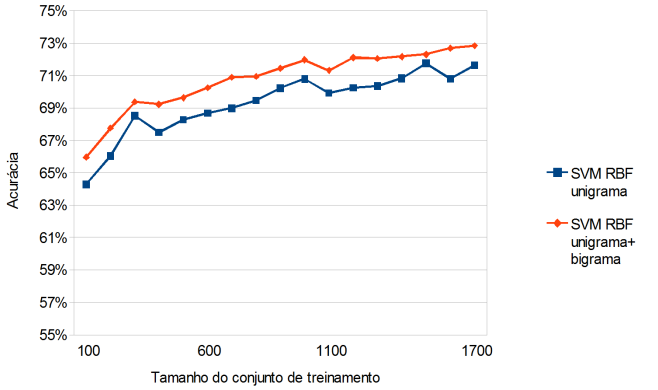

Fig. 5. Comparação entre a configuração unigramas VS unigramas+bigramas usando SVM com *Kernel* RBF como algoritmo de classificação

- Identificar textos que contém ou não *bullying* se mostrou uma tarefa difícil até mesmo para os anotadores. A subjetividade e ambiguidade se mostraram barreiras, e o *dataset* anotado não apresentou grande concordância entre os anotadores, mesmo provendo um manual de anotação com diretrizes e uma definição do que deveria ser considerado traço de *bullying*.
- Não foram feitos pré-processamentos nos dados para eliminar gírias e corrigir palavras escritas incorretamente.Estes ruídos podem ter prejudicado o desempenho dos classificadores.
- Apesar de termos selecionado apenas *tweets* cuja língua informada era o Português do Brasil, *tweets* erroneamente marcados como Português existem no córpus. Estes e outros ruídos podem ter influenciado negativamente na construção dos modelos.
- Informação importante para a tarefa pode estar contida em *tweets* pequenos, portanto a decisão de usar apenas *tweets* com mais de 120 caracteres pode ter influenciado negativamente os modelos construídos.
- Apenas algoritmos e implementações preexistentes foram utilizadas. O emprego de algorítimos novos ou implementações customizadas para a tarefa pode gerar melhores resultados.
- Como mencionado anteriormente, a curva de aprendizado continua crescendo, o que significa que um maior conjunto de dados anotados pode aumentar a acurácia.

B. Fase 2 - Classificar o papel dos envolvidos no episódio de bullying

Na Fase 2, utilizando unigramas e bigramas, SVM com *kernel* RBF, SVM com *kernel* Linear e Regressão Logística tiveram desempenho similar, apresentando 79.54%, 79.42% e 78.92% de acurácia respectivamente. A acurácia de 70.91% foi obtida pelo classificador Naive Bayes .

Para esta tarefa, SVM com *kernel* RBF obteve a Precisão de 60%, *Recall* de 58% e F-score de 56%. De forma a

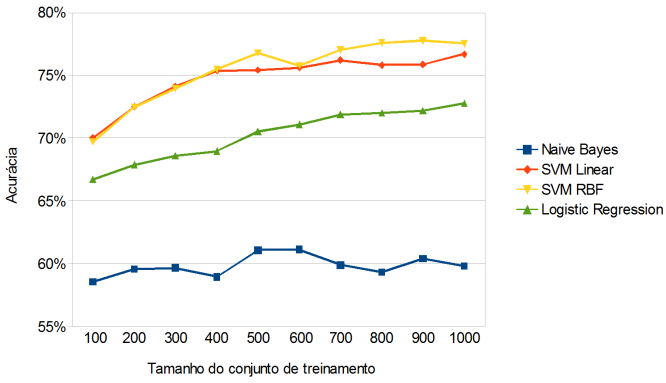

Fig. 6. Resultados dos experimentos realizados na Fase 2 - Classificar o papel dos envolvidos no episódio de *bullying* (Ver Figura 2) utilizando somente unigramas

melhor ilustrar os resultados, a Tabela IV exibe a matriz de confusão extraída do classificador com a maior acurácia. A matriz de confusão trás a informação de que os *tweets*, foram classificados segundo este modelo considerando somente duas classes – Vítima e Relator, ou seja a classe *Bully* não foi obtida como resultado em nenhuma das classes previstas. É válido ressaltar que o córpus dispõe de apenas 102 *tweets* rotulados como *Bully*, e ao que tudo indica, mesmo que uma nova anotação fosse realizada para aumentar o número de amostras, tal classe continuaria sendo a minoritária.

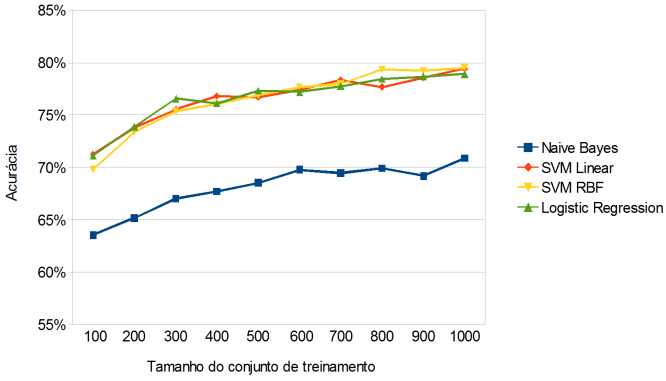

Fig. 7. Resultados dos experimentos da Fase 2 - Classificar o papel dos envolvidos no episódio de *bullying* (Ver Figura 2) utilizando unigramas e bigramas

A classe *bully* corresponde a apenas 8.7% do conjunto de dados utilizado. O problema de classificação quando as classes possuem instâncias com pouca representatividade devido ao custo da anotação humana são bem discutinos nas diversas tarefas da área de Processamento de Linguagem Natural, especialmente na língua portuguesa. Neste trabalho, os resultados obtidos na segunda fase foram prejudicados pela baixa quantidade *tweets* considerados frutos de desabafos de vítimas e de praticantes de *bullying* (ver Figuras 6,7 e Tabela IV).

TABELA IV
MATRIZ DE CONFUSÃO UTILIZANDO SVM RBF COM ATRIBUTOS COMPOSTOS DE UNIGRAMAS E BIGRAMAS PARA OS DADOS DE TESTE - 173 *tweets* (FASE 2 - CLASSIFICAR O PAPEL DOS ENVOLVIDOS NO EPISÓDIO DE *bullying*)

		Classe real		
		Bully	Vítima	Relator
Classe prevista	Bully	0	0	0
	Vítima	9	87	6
	Relator	6	14	51

VII. CONSIDERAÇÕES FINAIS

Bullying é um problema mundial, sendo que a agressão física ou moral repetitiva deixa sequelas psicológicas na pessoa atingida. A grande quantidade de dados disponíveis no Twitter é um dos aspectos motivadores para o presente trabalho, além do que a compreensão das ferramentas virtuais de comunicação pode levar a um entendimento maior do *bullying* fora da Internet. O estudo de detecção de traços de *bullying* em redes sociais, como o Twitter, para a língua inglesa tem tido grande ajuda do aprendizado de máquina e mineração de dados, uma vez que tais técnicas ajudam a obter informações sobre o comportamento dos indivíduos nestes ambientes, fornecendo subsídios para o combate e prevenção de tais atos.

Neste trabalho comparamos vários classificadores e conjuntos de atributos de forma a identificar qual melhor se adequa a tarefa de detecção automática de *Bullying* no Twitter na língua portuguesa. Utilizando uma base de dados de 2.000 *tweets* anotada manualmente, foram testados 4 classificadores e duas representações de atributos inicialmente, sendo o SVM com *kernel* RBF o que obteve os melhores resultados, com acurácia de 72.8%.

A inferência do papel do autor do *tweet* no episódio (os papéis de vítima, *bully* e relator foram anotados no *córpus* e são definidos na seção II-A) se mostrou uma tarefa complexa que demandou a inserção de novos atributos para ampliar o espaço de aprendizado. Neste trabalho, os resultados obtidos na previsão de tais papéis foram prejudicados pela baixa quantidade *tweets* nos quais o autor era o praticante de *bullying*, sendo necessário incluir novos *tweets* para estas classes. Em trabalhos futuros, técnicas mais sofisticadas de aprendizado de máquina podem ser testadas nesta tarefa, como *Ensemble Learning* [26] e *Deep Learning* [27]. Além disso, podem ser testadas técnicas de aprendizado não supervisionado para detectar as palavras mais frequentes em relatos de *bullying*, e usar a incidência destas palavras como um atributo para a tarefa.

REFERÊNCIAS

[1] D. Olweus, "Bullying at school," in *Aggressive behavior*, Springer, 1994, pp. 97–130.

[2] K. Nylund, A. Bellmore, A. Nishina, and S. Graham, "Subtypes, severity, and structural stability of peer victimization: What does latent class analysis say?" *Child development*, vol. 78, no. 6, pp. 1706–1722, 2007.

[3] J. Archer and S. M. Coyne, "An integrated review of indirect, relational, and social aggression," *Pers Soc Psychol Rev*, vol. 9, no. 3, pp. 212–230, 2005.

[4] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?" *Scandinavian journal of psychology*, vol. 49, no. 2, pp. 147–154, 2008.

[5] K. Rigby, "Health consequences of bullying and its prevention," *Peer harassment in school: The plight of the vulnerable and victimized*, p. 310, 2001.

[6] M. R. Leary, R. M. Kowalski, L. Smith, and S. Phillips, "Teasing, rejection, and violence: Case studies of the school shootings," *Aggressive behavior*, vol. 29, no. 3, pp. 202–214, 2003.

[7] UOL Notícias, *Autor do massacre no rio sofreu bullying, dizem ex-colegas de escola*, http://noticias.uol.com.br/cotidiano/ultimas-noticias/2011/04/08/autor-do-massacre-no-rio-sofreu-bullying-dizem-ex-colegas-de-escola.htm Acessado em 16-09-2017, 2011.

[8] Folha de SP, *Segundo a polícia, atirador de taiúva escolheu alvos*, http://www1.folha.uol.com.br/folha/cotidiano/ult95u67698.shtml, Acessado em 16-09-2017, 2003.

[9] D. C. Malta, M. A. I. Silva, F. C. M. d. Mello, R. A. Monteiro, L. M. V. Sardinha, C. Crespo, M. G. O. d. Carvalho, M. M. A. d. Silva, and D. L. Porto, "Bullying nas escolas brasileiras: Resultados da pesquisa nacional de saúde do escolar (pense), 2009," *Ciência & Saúde Coletiva*, vol. 15, no. suppl 2, pp. 3065–3076, 2010.

[10] Presidência da República, *Lei nº 13.185, de 6 de novembro de 2015.* http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2015/Lei/L13185.htm, Acessado em 16-09-2017, 2015.

[11] F. Resnik, A. Bellmore, J. Zhu, and W. Zhang, "Using machine learning to understand changes in how youth discuss bullying with celebrities on social media," in *Proceedings of the Technology, Mind, and Society*, ACM, 2018, p. 34.

[12] S. P. Murali, "Detecting cyber bullies on twitter using machine learning techniques," *Int'l J. Info. Sec. & Cybercrime*, vol. 6, p. 63, 2017.

[13] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, Association for Computational Linguistics, 2012, pp. 656–666.

[14] J.-M. Xu, X. Zhu, and A. Bellmore, "Fast learning for sentiment analysis on bullying," in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, ser. WISDOM '12, Beijing, China: ACM, 2012, 10:1–10:6, ISBN: 978-1-4503-1543-2.

[15] J. J. da Silveira Marciano, E. M. A. M. Mendes, and M. F. S. Barroso, "Cyberbullying classification using extreme learning machine applied to portuguese language," in *Latin American Workshop on Computational Neuroscience*, Springer, 2017, pp. 109–117.

[16] P. C. T. Fortuna, "Automatic detection of hate speech in text: An overview of the topic and dataset annotation with hierarchical classes," 2017.

[17] C. Salmivalli, "Participant role approach to school bullying: Implications for interventions," *Journal of adolescence*, vol. 22, no. 4, pp. 453–459, 1999.

[18] R. José *et al.*, "Estudo da ocorrência de cyberbullying contra professores na rede social twitter por meio de um algoritmo de classificação bayesiano," 2012.

[19] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining, (first edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005, ISBN: 0321321367.

[20] A. Bellmore, A. J. Calvin, J.-M. Xu, and X. Zhu, "The five w's of "bullying" on twitter: Who, what, why, where, and when," *Computers in human behavior*, vol. 44, pp. 305–314, 2015.

[21] K. Nalini and L. J. Sheela, "Classification of tweets using text classifier to detect cyber bullying," in *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2*, Springer, 2015, pp. 637–645.

[22] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on twitter," *ArXiv preprint arXiv:1702.06877*, 2017.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, ISSN: 0885-6125. DOI: 10.1023/A:1022627411411. [Online]. Available: https://doi.org/10.1023/A:1022627411411.

[24] C.-W. Hsu, C.-C. Chang, C.-J. Lin, *et al.*, "A practical guide to support vector classification," 2003.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15.

[27] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research [review article]," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, May 2014, ISSN: 1556-603X.