O uso de características estilométricas para identificação de autoria em mensagens curtas de origem multimodal

Rosaine Fiorio[1], Rauany J. Esperandim[1], Paulo Junior Varela[1] e Jhonnatan Ricardo Semler[1]

[1]Licenciatura em Informática – Universidade Tecnológica Federal do Paraná (UTFPR) – Francisco Beltrão-PR – Brazil

rosainefiorio@gmail.com, raulesperandim@gmail,com
paulovarela@utfpr.edu.br, jhonnatanricardo@gmail.com

Abstract. *The areas of identification of authorship focus primarily on medium to large texts. Therefore this work has as objective to present the characteristics of multimodal language can help identify the authorship of text messages and so spread this practice*

Resumo. *As áreas de identificação de autoria se concentram, principalmente em textos médios a grandes. Sendo assim esse trabalho traz como objetivo apresentar características da linguagem multimodal que possam auxiliar na identificação de autoria em mensagens curtas e assim difundir essa prática.*

1. Introdução

As áreas em que a identificação de autoria tem se concentrado geralmente são em textos considerados médios e grandes, tais como: textos literários, colunas de jornais, peças teatrais entre outros. Por outro lado, esses estudos são menos vistos em mensagens curtas, que muitas vezes são advindas principalmente de redes sociais e dispositivos móveis. O objetivo desse artigo é apresentar as características da linguagem multimodal que possam auxiliar peritos e linguistas na resolução de casos de identificação de autoria em mensagens curtas através do uso da estilometria.

2. Estilometria

A linguagem humana é a combinação de sons significativos para produzir a linguagem natural. Cada pessoa possui preferências e peculiaridades ao se expressar pelas diversas formas de linguagens, onde temos como principais: a escrita e a fala. São essas diferenças que compõem o estilo de cada autor [Varela *et. al*. 2011].

De acordo com [McMenamin 2002] o estilo é um reflexo de grupo ou a variação individual na linguagem escrita e se faz através de escolhas do escritor. Essas escolhas representam variações dentro da norma vigente como as diferentes formas de se dizer a mesma coisa, desvios dessa norma e as idiossincrasias. Para tanto, o estilo do autor transparece no conjunto dos padrões gramaticais que é geralmente resultado do uso de repetições de algumas ou de todas as formas do conjunto construído.

3. Linguagem Multimodal

As mudanças ocorridas devido à tecnologia da informação, principalmente pelo advento da internet, correram paralelas às mudanças na ortografia da língua portuguesa surgindo uma nova forma de comunicação, que [Hammes 2010] chama de “internetês”. A partir dessas mudanças uma terceira forma de comunicação seria gerada com características

próprias e marcantes em relação às duas existentes: verbal e não verbal. Ela integra som, imagem, texto e animação e seria chamada “linguagem multimodal” [Demo 2008]. Nesse trabalho é somente tratado as características multimodais em forma de textos e símbolos, como as palavras, abreviações e simbologias criadas à partir de pontuações.

Os usuários da internet compartilham de características especificas para se comunicar que segundo [Duarte 2009] são: uso de abreviações para acelerar o processo da escrita; troca de letras para tornar mais próxima do som da fala; repetições e interjeições que conferem um tom oral e expressam sentimentos do orador; falta de acentuação para agilizar a escrita, e, além disso, se sustenta devido à configuração dos teclados e que se realizam em trocas da letras acentuadas; utilização de letras maiúsculas para demonstrar sentimentos de agressividade; utilização de ícones multimídias, conhecidos como “smiles”, que são caretinhas que ilustram o estado de espirito do comunicador e que podem ser dos mais diversos tipos; e, invenções que no âmbito da internet são uma sequência desconexa de letras que caracterizam uma “risada internética”

De acordo com [Orebaugh e Allnutt 2009], a atribuição de autoria em mensagens curtas utiliza um conjunto de características estilométricas que permanece constante para um número alto dessas mensagens. Essas características definem o estilo de um autor, permitindo assim sua identificação.

4. Metodologia e Base de Dados

Para o desenvolvimento dos experimentos foram coletados 240 postagens de diferentes autores que possuem conta no Facebook. Além disso respeitou-se o limite de 20 palavras para cada texto coletado. Para tanto, trabalhar com pouco informação e gerar dados confiáveis consiste em um desafio.

A obtenção de resultados satisfatórios está em uma das etapas mais importantes que é a escolha dos atributos estilométricos que possam ser discriminatórios dentro do contexto em que a análise ocorre. No entanto, para realizar a extração dessas características dentro de textos curtos e com tantas peculiaridades como os que são objetos desse estudo foi necessário convergir em duas correntes: a sintática e a estrutural, respeitando as particularidades citadas por [Orebaugh e Allnutt 2009], e por Duarte (2009), para poder alcançar um alto índice de confiabilidade nos resultados**.**

Na tabela 1, podemos visualizar algumas das características estilométricas utilizados neste trabalho.

Tabela 1 – Características Estilométricas - Adaptado de [Orebaugh e Allnutt 2009]

Características Estilométricas	
Características	**Exemplo**
Frequência de caracteres (maiúsculas / minúsculas, números e caracteres especiais)	@ # $ % ^ & * - _ + = ‘ \
Frequência de símbolos (*Emoticons*)	:-) :) :-(:(;-) ;) :-P :P ;-P ;P :-D :D :’-(:’(:* :-*
Saudações	Oi, ola, bom dia, boa noite, boa tarde
Frequência de abreviaturas	Ta, q, vcs, abs, bjs, sdd, pq, d, gnt, tds, susse
Frequência de pontuação	. , ! ?

Média de palavras por frase	5 , 8.
Erros de ortografia	Palavras escritas de forma errada. Ex: complada.
Despedida	Bye, tchau, xau
Frequência de palavras função	Verbos, pronomes, e outras classes gramaticais.

O método aplicado neste trabalho consiste primeiramente na coleta da base de dados e seu pré-processamento (organização em formato texto – *ASCII*). Após, ocorreu a extração das características, utilizando-se da frequência relativa das características estilométricas apresentadas na Tabela 1. De posse das frequências das características, estas foram categorizadas conforme sua classe (saudações, abreviaturas, pontuação, *emoticons*, entre outros), que foram armazenados através de vetores. Ao final, são realizados os experimentos através de diversas abordagens de classificadores, para ver quais são as melhores técnicas para uso em mensagens curtas e que utilizem símbolos e textos de origem multimodal. As amostras dos textos foram divididas em 2/3 para treinamento e 1/3 para testes.

5. Conclusão

O desenvolvimento de trabalhos que abordem características multimodais através de textos em mensagens curtas ainda é recente e embrionário no Brasil. Sendo assim, consiste em um assunto a ser explorado nas suas mais diversas abordagens. Este trabalho relatou algumas das características estilométricas que podem ser utilizadas quando surgirem casos em que podem ser aplicados estes métodos para atribuição de autoria.

Referências

Demo, P.(2008) “Os desafios da linguagem do século XXI para o aprendizado na escola”. *Palestra, Faculdade OPET*, junho 2008. Diponivel em: http://www.nota10.com.br. Acesso em 14/10/2013

Duarte, L.(2009) “O jeito de escrever da internet invadiu a sala de aula”. In: *Porto Alegre: Jornal Zero Hora*.

Hammes, M H.(2010) “As mudanças que as novas tecnologias da escrita ofertadas pelo computador e pela Internet imprimem no meio educativo”. In: *Revista Digital* - Buenos Aires - Año 15 - Nº 145 - Junio de 2010.

Mcmenamin, G R.(2002) “Forensic Linguistics - Advances in Forensic Stylistics”. *CRC Press*, Florida-USA, 1a edition.

Orebaugh, A.; Allnutt, J. (2009). **“**Classification of Instant Messaging Communications for Forensics Analysis”. In: *The International Journal of FORENSIC COMPUTER SCIENCE*. Volume 1, pg 22-28.

Varela, P. J. Justino, E. J. R. Oliveira, L. E.S.(2011) “Identificação de Autoria de textos através do uso de Classes Linguísticas da Língua Portuguesa”. 8Th Brazilian Symposium in Information and human Language Technology - STIL 2011.Cuiabá, Brasil.