CENTRO UNIVERSITÁRIO UNA

DIRETORIA DE EDUCAÇÃO CONTINUADA, PESQUISA E EXTENSÃO

CURSO DE PÓS-GRADUAÇÃO ENGENHARIA DE SOFTWARE

Classificação de fluxo de sentimentos e a problemática do *Sentiment Drift*

Aluno: Fabrício dos Reis Neto Guimarães

Professor Orientador: Yóris Linhares de Souza

Belo Horizonte/MG

2013/01

Sumário

Resumo 3

Tema 4

Justificativa 4

Problematização 4

Objetivo 4

Referencial Teórico 5

Procedimentos Metodológicos 5

Categorização Hierárquica através de reconhecimento de emoções 6

Concept Drift e Sentiment Drift 9

DDM – *Drift Detection Method* 10

Analise de fluxo de informações utilizando janela dinâmica 11

Considerações finais e trabalho futuro 12

Bibliografia 13

Resumo

A análise das opiniões pessoais de usuários na web é chamada de análise de sentimento. Esta análise classifica o sentimento em neutro, positivo ou negativo em relação ao fluxo de informações que está em análise no momento. Com o passar do tempo há uma mudança de contexto no tópico, mudando o tema ou as pessoas mudando de opinião. Esta mudança é chamada de *concept drift* e ela pode ocorrer sutilmente com o passar do tempo ou abruptamente. A detecção deste *concept drift* é uma tarefa que esta sujeita a erros, como na análise das palavras no fluxo, e também na detecção errônea da mudança (*noise* ou ruído). Este trabalho abordará técnicas de *concept drift* no trabalho de [6] e estudos de análise de sentimentos [1][4].

Palavras-chave: mineração de dados, análise de sentimento, *concept drift*

Tema

A análise de informações em redes sociais, blogs, fóruns, vindas de pessoas é chamada de análise de sentimento. Refinando estas informações, retiram-se informações dos sentimentos das pessoas em relação ao que esta em destaque, como por exemplo, analisar tendências de votos em eleições, debates sobre atletas e personalidades e tendências de mercado.

Essas informações são um fluxo constante e com o passar do tempo há uma mudança de conceito gradual ou abrupta no contexto em análise. Essa mudança é chamada *concept drift* e aplicando-a para análise de sentimentos, tem-se o *sentiment drift*. Esta mudança pode ocorrer por novos eventos, manipulação da informação, falta de informação ou mudança de perspectiva [5].

Justificativa

Através da análise de dados anteriores, é possível prever tendências futuras de comportamento. Quando há uma mudança de tópico, ela deve ser detectada para modificar os dados que estão sendo analisados.

Problematização

Na análise de informações vindas de pessoas, há o problema de possuir palavras com sentidos diferentes, sarcasmos, palavras escritas erradas, chamado de ruído.

A detecção de *concept drift* esta sujeita a erros, tanto na análise de dados errados como no treino do algoritmo.

Objetivo

O objetivo deste trabalho é iniciar um estudo sobre o tema de mineração de informações mais especificamente a análise de sentimentos. O estudo de *concept drift* e aplicado a isto, o *sentiment drift*. Será descrita algumas abordagens sobre

classificação de sentimento e a mudança de tópico (*drift*) e as conclusões que seus autores chegaram.

Neste estudo inicial, serão analisadas técnicas destas abordagens e uma conclusão sobre estes estudos para a criação de um trabalho futuro.

Referencial Teórico

Os estudos abordados e que serão descritos, são [5] e [6] que abordam o conceito de *concept drift* e *sentiment drift* e algumas técnicas. Os estudos [1] e [4] abordam, o conceito de categorização de dados em sentimentos (positivo, negativo ou neutro) [1] e uma técnica de análise de sentimentos utilizando o conceito de *sentiment drift* [4].

Outros estudos deram embasamento básico para o entendimento destes conceitos e podem ser encontrados na referência bibliográfica.

Procedimentos Metodológicos

Deste estudo, foi feita uma revisão sobre conceitos, demonstrando um estudo inicial sobre o assunto. Foram estudadas diversas abordagens na grande área mineração de dados, e entrando em conceitos mais específicos como análise de sentimentos, mudança de tópico (*concept drift*) e a mudança de sentimento (*sentiment drift).*

Tipo de pesquisa foi um estudo sobre casos e metodologias.

Categorização Hierárquica através de reconhecimento de emoções

Categorização hierárquica lida com os problemas de categorização onde as categorias são organizadas em hierarquias. A categorização dos textos em forma de categorização hierárquica tem sido utilizada por nós humanos na organização do nosso conhecimento [1].

Este método de organização é dividido em dois tipos de classificação: o primeiro, *two-level hierarchy* onde o texto é representado na relação de neutralidade e emoção. O segundo, *three-level hierarchy,* que além de separar o texto em neutro e emotivo, o emotivo ainda é dividido por polaridade, onde possui seis tipos de polaridades, sendo uma positiva (*happiness*) e cinco negativas (*sadness, fear, surprise, disgust, anger*).

Two-level hierarchy

A principal tarefa desta classificação é descobrir o quanto a presença de palavras neutras no texto afeta o desempenho na quebra do texto em classes que contenham emoção ou não [1].

No primeiro nível as classes de texto são separadas em classes emotivas e classes não emotivas. Com isto, a maior parte das classes são encaixadas como não emotivas, fazendo com que a classificação das classes emotivas mais precisa, pois separou-se classes em que a emoção não é parte do texto.

Após esta primeira separação, a segunda classificação é feita em duas etapas: uma considerando que o primeiro nível, as classes não emotivas, esta separada corretamente, sendo necessário apenas preocupar-se somente com os erros encontrados neste segundo nível (chamado de *golden standard*). O segundo experimento leva em consideração que o primeiro nível possui erros e com isto, a classificação é feita sem levar em conta as classes não emotivas.

Quando comparado os resultados do *golden standard* com o segundo experimento, verificou-se que a precisão nos dados do *golden standard* possuem uma maior precisão no que diz respeito as classes emotivas negativas, pois não considera

as classes não emotivas. No segundo experimento, desconsiderando as classes não emotivas, a classe emotiva positiva, ficou com maior parte da classificação, mostrando um maior balanceamento entre as classes negativas, pois somente as classes que possuíam emoção estavam em análise. Na classificação *three-level hierarchy* é esperado que a precisão das classes negativas seja maior, pois separa as classes emotivas das não emotivas.

Three-level hierarchy

Nesta fase, [1] dividiu as 7 classes de classificação em 3 níveis, sem emoção, felicidade(*happiness*), e as outras 5 de sentimento negativos.

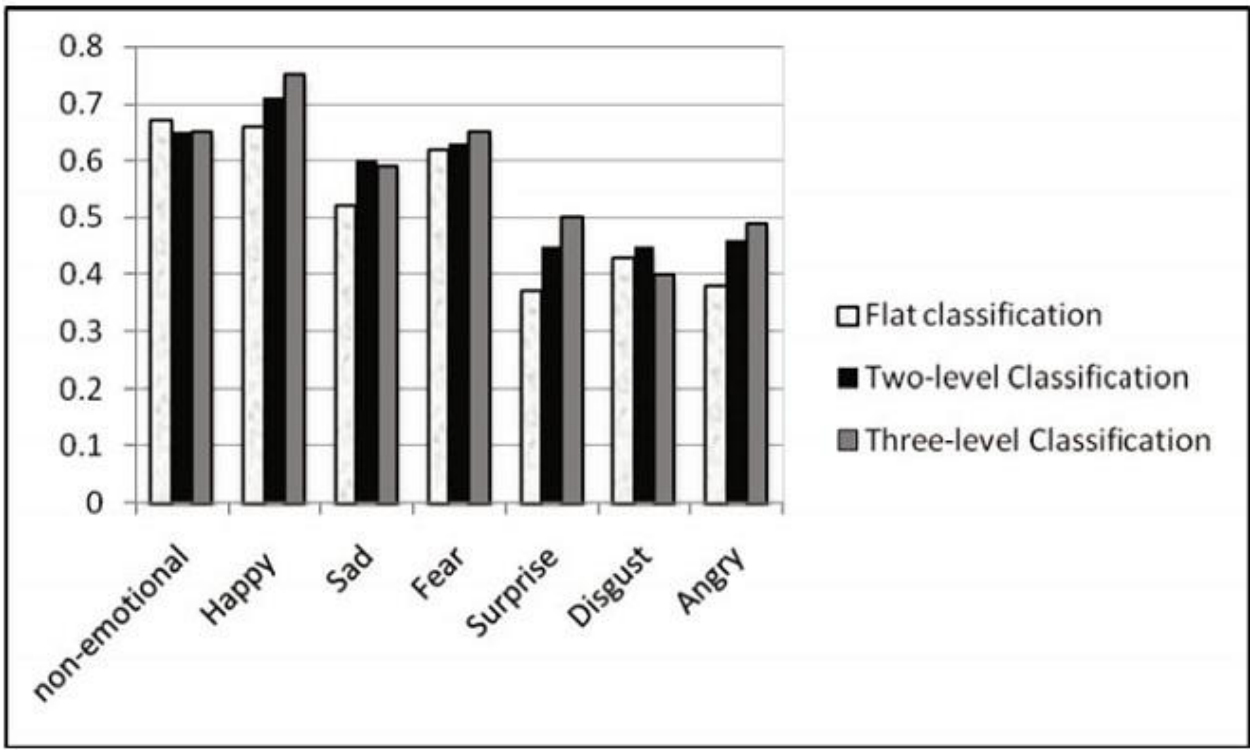

Figura 1: Comparativo entre as classificações [1]

Na classificação em três níveis, o primeiro nível determina a separação de classes emotivas das não-emotivas. Após isto, no segundo nível, separa-se as classes com emoção positiva das classes com emoção negativa. Após isto, no terceiro nível, classifica-se as classes com emoção negativa em *sad, fear, surprise, disgust, angry.*

Com isto observou-se que na classificação em dois níveis, os dados ficaram quase balanceados entre classes emotivas e não-emotivas, e na classificação de três níveis, aumentou-se o balanceamento entre todas as classes e com isso aumentou a precisão de acerto.

Com isso concluiu-se que na classificação por emoções, possuir uma classe não-emotiva decresce os resultados significativamente, mas na classificação *two-level hierarchy*, que retira a classe não-emotiva, o resultado teve uma melhor classificação.

Na segunda parte, onde foi a classe emotiva positiva *happiness* e aplicado a classificação *three-level hiearchy*, os resultados foram ainda mais balanceados.

Um problema visto foi que, quando classificando grande massa de dados desbalanceadas e não treinadas, os resultados são muito esparsos, pois quando o dado não se encaixa em nenhuma das classes, este dado vai para a classe onde possui mais instancias. Então é necessário um treinamento e um balanceamento de dados para utilizar esta metodologia.

Concept Drift e Sentiment Drift

Concept drift é a ocorrência de mudanças de contexto no tópico em estudo, considerando que novos tópicos sendo adicionados e removidos. Como estudado em [10], existem dois tipos de mudança (1) abrupta (*sudden*), como exemplo, mudança climática ou (2)gradualmente(*gradual*) que envolve o fator tempo, como exemplo as intenções de voto na política, como visto em [6]. A detecção automática desta mudança esta sujeita a problemas diversos como o tópico nem sempre esta evidente, distinguir entre um real *drift* de um *noise*, ruído (uma falsa mudança de tópico).

Existem técnicas esta detecção, onde podem ser divididas em: (1) seleção baseada em janela; (2) seleção baseada em peso; visto em [10], onde foi omitido (3) *ensemble learning*.

O conceito mais utilizado é a seleção baseada por janela. Uma janela é um conjunto finito de palavras, é selecionada e analisada. Com a chegada de mais palavras, estas são pré-processadas e verifica se ela é incluída ou não na janela de tópicos analisadas. Esta pré-processamento pode incluir retirar redundância, irrelevância e ruídos. Esta janela pode possuir tamanho fixo ou dinâmico, isto é, aumentando e diminuindo de tamanho acordo com alguma técnica de *concept drift*. Alguns algoritmos propostos são FLORA (Widmer and Kubat, 1996), FRANN (Kubat and Widmer, 1994) e TMF - *Time Windowed Forgetting* (Salganicoff, 1997) [10].

Na seleção baseada em peso utiliza-se de um processo de adicionar ou remover palavras de acordo com o peso associados a ela. Este peso leva em conta o tempo que a palavra esta em análise e o quão bem ela descreve o tópico.

Sentiment drift é o conceito de *concept drift* aplicado a palavras que envolvem algum tipo de sentimento. A palavra, ou conjunto de palavras é classificada de acordo com uma análise da léxica, e descobre-se o tipo de sentimento associado a esta, positivo, negativo e neutro. Um exemplo de algoritmo será descrito no próximo capítulo.

DDM – *Drift Detection Method*

No trabalho de [6], foi proposta uma alteração ao sistema de [10] *Drift Detection Methods*, os quais neste trabalho não serão demonstrados.

DDM utiliza uma contagem, durante a fase de pré-processamento, que leva em consideração o numero de erros produzidos pelo algoritmo durante seu treinamento. Esta técnica foi provada ser eficiente para casos em que as mudanças são abruptas. No caso de mudanças graduais, os exemplos que compõe o estudo podem crescer exponencialmente.

EDDM foi proposto para se levar em consideração mudanças graduais no fluxo de informações. Ela é feita levando em consideração a média da soma da distância de dois erros. Para a avaliação desta técnica, [6] utilizou três algoritmos de detecção de *concept drift*: uma árvore de decisão DNF e dois conhecidos como *nearest-neightboorhood*, que não serão abordados. A base de informações utilizadas para a avaliação possui um grade volume de dados, representa vários tópicos reais e apresenta características como *drift* abrupto e gradual, presença e ausência de ruídos, *stopwords* e números.

Analisando DDM x EDDM, [6] concluiu que, para mudanças abruptas de comportamento, ambas técnicas detectam esta mudança similarmente. Para mudanças graduais lentas, a técnica EDDM detectou as mudanças antes e mais vezes que o DDM.

Analise de fluxo de informações utilizando janela dinâmica

Como descrito anteriormente, uma janela dinâmica é utilizada para analisar o *concept drift*. O trabalho de [4], utiliza uma janela dinâmica onde o tamanho desta janela seja pequeno o suficiente para que não sofra influência de *sentiment drift* e seja grande o suficiente para que o algoritmo possa aprender com eficácia. Segundo Silva, I. S., Barbosa [4], seu trabalho faz manter um conjunto de treino que provê maior ganho de informação com um viés temporal objetivando descrever melhor o fluxo de sentimento atual. Além disso, não é necessário um pré-processamento para verificar a ocorrência ou não de *sentiment drift* fazendo com que não seja gasto uma quantidade considerável de tempo e recurso processamento.

Para se adicionar ou remover informações nesta janela, [4] propôs utilizar a função de *rank* onde a cada nova palavra eleita para ser inserida na janela outras que já estão na janela são eleitas para serem removidas. O algoritmo ira remover alguma palavra da janela que possui maior similaridade com a palavra que esta sendo inserido no treino. Assim, com este algoritmo de *rank*, as novas palavras que entrarem na janela de treino terão menor probabilidade de serem removidas quando outras mais novos forem analisadas.

Considerações finais e trabalho futuro

Neste trabalho foi feito uma abordagem de técnicas que compõe o estudo de mineração de informação. Foi introduzido o conceito e técnicas de detecção de *concept drift*, que deve ser levado em conta num fluxo contínuo de informações. Uma apresentação de como podem ser feitas as classificações de palavras, em atributos léxicos. Essa classificação é chamada de análise de sentimentos e, aplicado o conceito de *drift* nesta coleção, é chamado de *sentiment drift*.

Para a detecção do *concept drift*, como foi visto a presença de ruídos (*noise*), o tipo de *drift*, abrupto ou gradual, devem ser considerados.

Para a avaliação destas técnicas, é necessário um estudo em um conjunto de dados com informações genéricas, sobre diversos tópicos. Um problema geral encontrado foi a falta de massa de dados que refletem o mundo real para o treino dos algoritmos, com características diferentes e mudanças no fluxo de informações distintas.

Um futuro trabalho será um estudo baseado em janelas para a detecção de *sentiment drift* em fluxos constantes de informações, como exemplo, redes sociais, reviews de filmes por exemplo, entre outras.

Bibliografia

[1] Ghazi, D., Inkpen, D., Szpakowicz, S.. *Hierarchical Approach to Emotion Recognition and Classification in Texts*. Em: 23rd *Canadian conference on Advances in Artificial Intelligence,* páginas 40 a 50.

[2] Pang, B., Lee, L.. *Opinion Mining and Sentiment Analysis*. *Computer Science Department, Cornell University, Ithaca, NY 14853, U.S.A. http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf.* Acessado em 17/02/2013.

[3] Pang, B., Lee, L.. *A sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*. Em: 42nd *Annual Meeting on Association for Computational Linguistics,* 2004.

[4] Silva, I. S., Barbosa, G. A. R., Veloso, A., Ferreira, R. and Meira, Jr. W. Análise Adaptativa de Fluxo de Sentimento Baseada em Janela Deslizante Ativa. SBBD[1], 2011. http://homepages.dcc.ufmg.br/~ismael.silva/papers/ShortPaperSbbd11.pdf Acessado em 17/02/2013.

[5] Durant, K. T.. *Sentiment Drift and Its Effect on the Classification of Web Log Posts*. 15th *ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 1275 a 1284, 2008.

[6] Baena-García, M., Campo-Ávila, J. d., Fidalgo, R., Bifet A., Gavaldà, R., Morales-Bueno, R.. *Early Drift Detection Method*. ICAIS'11 *Proceedings of the Second international conference on Adaptive and intelligent systems*, páginas 88 a 97, 2011.

[7] Lindstrom, P., Delany S. J., Namee, B. M.. *Handling Concept Drift in a Text Data Stream Constrained by High Labelling Cost*, *Florida Artificial Intelligence Research Society Conference* (FLAIRS), 2010.

[8] Pang, B., Lee, L., Vaithyanathan, S.. *Thumbs up? Sentiment Classification using Machine Learning Techniques*, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* – Volume 10, páginas 79 a 86, 2002.

[9] Tsymbal, A. *The problem of concept drift: definitions and related work*, *Computer Science Department, Trinity College Dublin*, 2004.

[10] Gama, J., Medas, P., Castillo, G., Rodrigues, P.. *Learning with drift detection. Advances in Artificial Intelligence–SBIA,* páginas 66 a 112, 2004.

[1]SBBD – Sociedade Brasileira de Banco de Dados