SAMH: a platform for Sentiment Analysis for Mental Health

Francisco Maciel ei11084@fe.up.pt
Hugo Sousa ei11083@fe.up.pt
Ricardo Silva ei11079@fe.up.pt

November 13, 2015

Abstract

Este trabalho surge da análise de dados de 2 projetos que pretendem analisar e descobrir métodos mais eficazes no tratamento da depressão, recorrendo a tecnologia. O projeto STOP DEPRESSION aglomera um conjunto de transcrições de consultas de psicologia; o E-COMPARED fornece um conjunto de ratings referentes a estados psicológicos de pacientes. Tendo também uma base lexical, que associa termos a estados emocionais, desenvolveu-se uma plataforma que permite ao utilizador uma pesquisa facilitada de documentos ou dados relativos aos pacientes, de forma a perceber o seu estado emocional.

1 Introdução

O *SAMH* é uma plataforma que permite a pesquisa customizada e facilitada de documentos, a partir de vários parâmetros definidos pelo utilizador. Esta pesquisa é efetuada sobre um *dataset* constituído por dados textuais relativos a consultas de vários pacientes, no âmbito da saúde mental.Os principais utilizadores alvo desta plataforma são profissionais da área, que através destas pesquisas podem inferir o estado mental dos pacientes. Os resultados recolhidos utilizando a plataforma podem ser comparados à analise psicológica efetuada pelo profissional, permitindo tirar novas ilações, resultando em estudos mais completos sobre estado do paciente.

Este projeto surge como uma oportunidade de estudar e aproveitar dados provenientes de dois projetos diferentes mas relacionados entre si, o projeto *E-COMPARED*[1] e o projeto *STOP DEPRESSION*[2]. O projeto *E-COMPARED* é uma iniciativa europeia com o objetivo de investigar e divulgar recomendações relativas ao tratamento "misto" da depressão. Para alcançar este objetivo, recolhe e compara conhecimento relativo ao tratamento tradicional da depressão e o tratamento com recurso à tecnologia. À semelhança do projeto europeu, o projeto *STOP DEPRESSION* tem o mesmo objetivo, procurando aplicar novas tecnologias ao tratamento da depressão e à prevenção do suicídio no *Plano Nacional de Saúde Mental* em Portugal. Estes projetos fornecem assim uma quantidade de dados relevante para ser analisada, servindo como base para este projeto.

A plataforma desenvolvida pode ser considerada no contexto de *Sentiment Analysis* ou *Opinion Mining*[3, 4] que consiste no processamento de linguagem natural para extrair significado e informação a partir da fonte em análise. Assim sendo, combinando os termos da base lexical com técnicas de *information retrieval* sobre as transcrições fornecidas, será possível analisar e perceber o estado psicológico do paciente em questão.

2 Datasets

O projeto assenta sobre os dados recolhidos de dois projetos relacionados com a saúde mental e depressão, *STOP DEPRESSION* e *E-COMPARED*.

2.1 Dados do *STOP DEPRESSION*

Fazem parte deste *dataset* as transcrições de várias consultas relativas a 23 pacientes distintos. Estando o projeto *STOP DEPRESSION* a ser implementado em Portugal, estas consultas estão na língua portuguesa. As transcrições são

fundamentais para o objetivo desta unidade curricular, por se tratarem de dados textuais, potenciando o uso das ferramentas de *information retrieval*. Estas transcrições estão anonimizadas para proteger a privacidade dos pacientes.

2.2 Dados do *E-COMPARED*

A partir deste projeto europeu, são fornecidas avaliações numéricas, relativas a um dado parâmetro emocional (*Mood, Activities, Enjoy, Esteem, Sleep, Worrying, Social*) para um dado paciente, em sequência e identificados a partir de um *timestamp*, ao longo de um dado intervalo de tempo. Foram apenas retirados estes dados relativos a 23 pacientes alemães, o mesmo número de pacientes dos quais se tem transcrições das consultas.

2.3 Base Lexical em Português

Para permitir a análise lexical dos elementos textuais, é utilizada uma base lexical de termos emocionais que associa várias palavras a determinados estados sentimentais, que são agregados em 4 níveis de abstração, tendo por base emoção positiva, negativa, não especificada, surpresa ou indiferença. Assim, a identificação dos termos referenciados, permitirá tirar conclusões sobre o estado emocional do paciente.

2.4 Combinação dos datasets

Para melhor cumprir os objetivos deste projeto, foram aglomerados os dados dos dois projetos, combinando o estado psicológico de um paciente do projeto *E-COMPARED* diretamente a um paciente aleatório do projeto *STOP DEPRESSION*. Apesar de esta combinação poder resultar em incoerência a nível de resultados numa análise contextual, torna o *dataset* resultante mais rico e para o qual passam a ser possíveis pesquisas mais completas e interessantes ao âmbito da unidade curricular. No futuro, o projeto *STOP DEPRESSION* irá também incluir avaliações numéricas dos parâmetros emocionais dos seus pacientes, mas sendo que de momento estes dados não foram ainda colecionados, é simulada a sua existência recorrendo a dados externos de pacientes de outro país.

2.5 Qualidade dos dados e das fontes de informação

Os *datasets* referidos provêm de fontes de informação fidedignas e de projetos/instituições de renome nacional e internacional.

2.5.1 *STOP DEPRESSION*

Os dados provenientes deste projeto têm como origem consultas realizadas no *ISMAI* e noutros estabelecimentos de saúde. Estas consultas são filmadas e é feita a gravação de som. As transcrições são posteriormente efetuadas por psicólogos ou auxiliares. Assim, estes dados estão sujeitos ao erro humano e às limitações e perda de informação de conversão de um diálogo para formato textual. Existem também limitações na compreensão do material sonoro pois por vezes o responsável pela transcrição assinala incompreensão de alguns termos (Ex:"C: (incompreensível)"). Apenas um *subset* das terapias dos pacientes está presente, podendo mesmo haver sessões em falta na sequência das transcrições. No entanto, os dados estão ordenados e contêm entre cinco a quinze sessões por paciente.

2.5.2 *E-COMPARED*

Os dados provenientes deste projeto correspondem a métricas recolhidas através de soluções tecnológicas. Foram escolhidos dados de pacientes alemães e são correspondentes a vários parâmetros emocionais. A incoerência do *subset* recebido verifica-se na falta de variedade nos parâmetros emocionais para um dado paciente, uma vez que dos sete parâmetros presentes, cada paciente pode ter informação apenas de um ou dois, limitando o potencial dos dados.

2.5.3 *Base Lexical*

Esta base lexical, designada *EMOTAIX.PT*[5] foi traduzida a partir de um projeto francês designado *EMOTAIX*[6]. É ainda composta por termos emocionais retirados de uma outra base lexical portuguesa mais genérica designada de *PORLEX*[7]. Para além disso, os termos lexicais apresentam pouca diferenciação entre género, número e tempo verbal, o que requer um investimento adicional da ferramenta para classificar termos que sejam semelhantes mas não idênticos aos termos da base lexical.

2.6 Caracterização e propriedades dos datasets

2.6.1 *STOP DEPRESSION*

Para cada paciente, existe um número variável de transcrições, não sendo necessariamente consultas consecutivas. Cada transcrição está identificada com a data, identificador do paciente e terapeuta. O formato destas transcrições segue um formato standard, em que as citações são antecedidas por um "C:" caso de trate do cliente/paciente ou por "T:" caso se trate do terapeuta.

Segundo esta estrutura a transcrição segue forma ilustrada na Figura 1.

T: trabalho não é? hm

C: como se nada tivesse acontecido

T: mm-hm

C: como se todo aquele rol d emoções que eu estava a sentir - - num não tivessem passado, nada e transmitia-lhe isso

Figure 1: *Excerto de uma transcrição*

2.6.2 *E-COMPARED*

Os dados fornecidos correspondem a 23 pacientes alemães, e indicam para um dado instante o *rating* de um dado estado emocional para um dado paciente. Este *rating* é um valor de 1 a 10 em intervalos de 0.1. Cada avaliação tem associado um *timestamp* já ordenado.

P004	898D4C1F	2015-04-10 09:09:31.000	sleep	4 unused comment 2015-04-10 08:10:59.640
P004	898D4C1F	2015-04-09 09:56:10.000	sleep	9 unused comment 2015-04-09 08:57:15.490
P004	898D4C1F	2015-04-08 09:22:39.000	sleep	7,5 unused comment 2015-04-08 10:15:46.313
P004	898D4C1F	2015-09-19 20:41:29.000	social	3 unused comment 2015-09-20 08:13:02.947
P004	898D4C1F	2015-09-18 20:51:49.000	social	8 unused comment 2015-09-18 19:54:18.247
P004	898D4C1F	2015-09-03 20:30:02.000	social	7 unused comment 2015-09-04 06:10:08.250

Figure 2: *Excerto de ratings dos estados sleep e social associados ao paciente P004*

2.6.3 *Base Lexical*

A base lexical pode ser vista e definida como uma árvore, estando dividida em quatro níveis de profundidade. Apresentando da raíz para as folhas, estes níveis podem ser designados de: primário, global, intermediário e específico. Abaixo deste último nível encontram-se os termos que se associam ao estado psicológico de uma pessoa, referidos nos níveis superiores. Na Figura 3 encontram-se (de cima para baixo) os níveis primário, global, intermediário e específico. A linha seguinte indica o número de termos, encontrando-se estes imediatamente abaixo.

POSITIVA	NEGATIVA
MAL ESTAR	BEM ESTAR
DEPRESSÃO	ALÍVIO
APATIA	APAZIGUAR
86	64
amorfo	acalmar
apatia	ajuda
apática	ajudar
apático	aliviar

Figure 3: *Excerto da base lexical em português*

3 Modelo conceptual do domínio do problema

Tendo em conta os *datasets* disponíveis, é possível identificar entidades e relações entre estas no domínio do problema. As consultas de psicologia são um diálogo entre o terapeuta e o paciente, que são mais tarde transcritas num documento. Estas transcrições contêm múltiplos termos, que podem estar incluídos na base lexical fornecida. Um termo não é necessariamente precedido dos 4 níveis na base lexical, podendo suceder diretamente o nível primário.

A um paciente estão também associados vários *ratings* de diferentes estados emocionais.

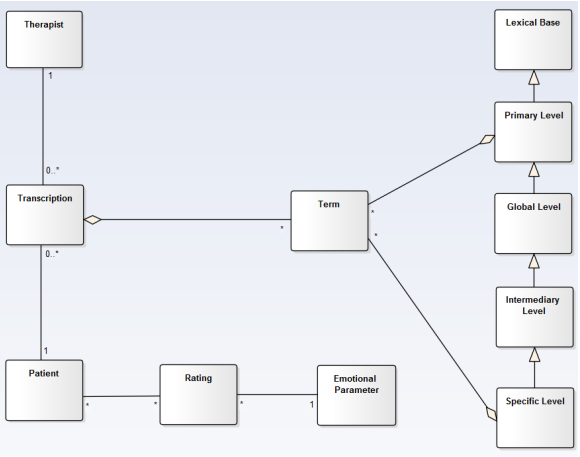

Figure 4: *Modelo conceptual*

4 Plataforma *SAMHealth*

A aplicação desenvolvida é uma ferramenta de pesquisa sobre os *datasets* referidos, que retorna os documentos mais relevantes para o utilizador, de acordo com os termos de pesquisa.

4.1 Funcionalidades - *information retrieval tasks*

A plataforma está dividida em dois módulos: pesquisa por termos e pesquisa pela base lexical. Tendo em conta um cenário de utilização real, é possível que os resultados sejam filtrados por um dado paciente, de forma a analisar o seu estado emocional, sem obter resultados não relevantes. Ambos os módulos retornam os documentos mais relevantes para a pesquisar efetuada, indicando alguns excertos com os termos encontrados, e permitem ainda o seu *download* para uma pesquisa mais aprofundada pelo utilizador. Retornam ainda informação do paciente, terapeuta, número e data da sessão.

4.1.1 Pesquisa por termos

Neste módulo, o utilizador insere os termos que pretende pesquisar nas transcrições. Por defeito, a *query* pretendida será um ou mais termos, que se pretendam encontras nas transcrições, separados por espaços. No entanto, a *query* executada é diretamente processada pela ferramenta de *information retrieval*. Assim sendo, tendo o utilizador conhecimento para tal, poderá executar *queries* mais complexas, podendo obter resultados que melhor correspondam às suas necessidades.

4.1.2 Pesquisa pela base lexical

Funcionando como um nível de abstração, para facilitar a pesquisa do estado emocional e justificando assim o *sentiment analysis*, é possível pesquisar por documentos através dos níveis da base lexical. A ferramenta é flexível, permitindo ao utilizador definir até ao nível de profundidade preferido. Internamente, a ferramenta trata de pesquisar os termos relativos ao nível especificado na base lexical e fazer uma pesquisa sobre esses.

4.2 Arquitetura

A plataforma trata-se de um website, que tem como suporte um servidor desenvolvido em *PHP* que serve as suas páginas, cuja arquitetura se esquematiza na Figura 5. Por sua vez, este servidor faz pedidos a um outro servidor, neste caso a ferramenta de *information retrieval*, onde se encontram armazenados os documentos e que retorna os resultados em resposta às *queries* efetuadas.

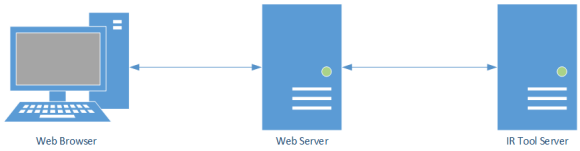

Figure 5: *Arquitetura da plataforma*

4.3 Ferramenta de *information retrieval*

Tendo em conta a grande quantidade textual presente nos *datasets* sobre o qual a plataforma

se baseia, recorrer a uma ferramenta de *information retrieval* foi uma decisão trivial.

Este tipo de ferramentas destinam-se a pesquisas textuais, tendo vantagens relativamente a bases de dados relacionais, permitindo pesquisas mais rápidas e sofisticadas. Por outro lado, a indexação (e consequentemente inserção) é mais lenta[8]. Neste projeto, este é um fator desprezável, visto os dados serem estáticos.

Para este projeto, foi escolhido o *Apache Solr*[9] como ferramenta de *information retrieval*, que se baseia no *Apache Lucene*[10]. Esta ferramenta é *user-friendly* e de fácil adaptação, providenciando um servidor com uma *API REST* de fácil interação.

O *Apache Solr* já aplica automaticamente conceitos importantes de *information retrieval*, tais como *tokenization*, *stemming* ou ignorar palavras comuns da linaguagem. Todos estes conceitos podem ser configurados para a língua portuguesa, linguagem usada nas transcrições.

4.3.1 Documento

Tratando-se de uma ferramenta de *information retrieval*, é importante definir o documento. Tendo em conta o domínio do problema, o documento refere-se a uma transcrição, ou seja, uma consulta. Desta forma, é possível recuperar informação relativa a uma consulta.

Dividir ainda mais o documento (assumir como documento cada fala, por exemplo), não daria resultados interessantes. Por um lado, algumas frases tornam-se demasiado pequenas ou sem significado e por outro, deixa de haver interação entre diferentes frases, que podem ser relevantes numa *query* textual.

Cada consulta corresponde a um documento físico em formato .doc ou .docx. De forma a extrair o conteúdo para o *Solr*, foi necessário recorrer à ferramenta *Apache Tika*[11].

4.3.2 Campos indexados

Para cada documento, os campos indexados são os seguintes:

- patient - identificador do paciente
- therapist - nome do terapeuta
- content - conteúdo da transcrição
- file - nome do ficheiro da transcrição
- session_number - número da sessão
- session_date - data da sessão

Atualmente, os campos imprescindíveis para o funcionamento da plataforma SAMH, são o *content* (de onde é pesquisado o conteúdo), o *patient* (para filtrar por paciente) e o *file* (para obter o caminho para o download do ficheiro).

4.3.3 Resultados

5 Conclusões

A plataforma demonstra ser bastante interessante e ter aplicabilidade prática num cenário real de utilização. É possível retirar informação emocional relevante das transcrições disponibilizadas.

As técnicas de pesquisa usadas no contexto da plataforma são, obviamente, rudimentares e simplistas no domínio de *sentiment analysis*, apesar de, de um modo geral, ser percetível o estado emocional de um paciente ao longo das sessões, pesquisando através dos níveis da base lexical.

Por exemplo, a negação de uma afirmação é algo complexo de identificar. Considerem-se os seguintes exemplos:

- 1 - "C: Eu não estou deprimido."
- 2 - "T: Está deprimido? / C: Não."
- 3 - "C: Deprimido nunca estive."
- 4 - "T: Não está deprimido? / C: Estou."
- 5 - "T: Está deprimido? / C: Não muito...mas já tive dias melhores."
- 6 - "T: Está deprimido? / C: Nem sabe quanto."

A plataforma detetará o facto de o paciente se encontrar deprimido em todas as afirmações anteriores. No entanto, pode verificar-se que nas afirmações 1, 2 e 3 tal não deveria acontecer, pois o paciente nega esse facto. Também é possível verificar nas afirmações 4, 5 e 6 que excluir este facto apenas pela proximidade de advérbios de negação não é suficiente e pode induzir em erro. Também o carácter subjetivo (por exemplo ironia, sarcasmo) das falas e das emoções são um fator limitativo desta análise. Esta é uma tarefa não trivial e que requer um maior investimento nesta área.

Relativamente à ferramenta de *information retrieval*, mostrou-se ser bastante poderosa e útil em determinadas aplicações como o caso da plataforma SAMH, onde a pesquisa textual é uma forte componente.

Como trabalho futuro, a plataforma pode ser ainda melhorada, tendo sido pensadas as seguintes funcionalidades:

- mostrar a evolução graficamente de um dado paciente ao longo do tempo para um dado nível da base lexical
- permitir o upload de novas transcrições
- associar os *ratings* dos pacientes e realizar *queries* sobre eles

References

[1] European Community's Seventh Framework Programme (FP7). E-COMPARED — European Comparative Effectiveness Research on Internet-based Depression Treatment. http://www.e-compared.eu/. Acedido em Outubro, 2015.

[2] Stop depression: Stepped care treatments and digital solutions for depression and suicide prevention in primary care. http://eeagrants.org/project-portal/project/PT06-0010. Acedido em Outubro, 2015.

[3] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, e Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics, Vol. 37, No. 2*, páginas 267–307, Junho 2011.

[4] Bo Pang e Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, Vol. 2, No 1-2*, página 1–135, 2008.

[5] Sara Costa, Rui A. Alves, Fernando Barbosa, e Thierry Olive. Sig writing porto. Em *SIG Writing Porto 2012*, número 13 em International Conference of the EARLI Special Interest Group on Writing. Universidade do Porto, Julho 2012.

[6] Inês Gomes e São Luís Castro. Porlex, a lexical database in european portuguese. *Psychologica, 32*, 2009.

[7] Annie Piolat e Rachid Bannour. An example of text analysis software (emotaix-tropes) use: The influence of anxiety on expressive writing. *Current psychology letters, Vol. 25, Issue 2*, páginas 91–108, 2003.

[8] Why Use Solr. https://wiki.apache.org/solr/WhyUseSolr. Acedido em Novembro, 2015.

[9] The Apache Software Foundation. Apache Solr. http://lucene.apache.org/solr/. Acedido em Novembro, 2015.

[10] The Apache Software Foundation. Apache Lucene. http://lucene.apache.org/. Acedido em Novembro, 2015.

[11] The Apache Software Foundation. Apache Tika - a content analysis toolkit. https://tika.apache.org/. Acedido em Novembro, 2015.