Sentiment Analysis for Mental Health

Francisco Maciel
ei11084@fe.up.pt

Hugo Sousa
ei11083@fe.up.pt

Ricardo Silva
ei11079@fe.up.pt

October 9, 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[1] aglomera um conjunto de transcrições de consultas de psicologia; o E-COMPARED[2] fornece um conjunto de ratings referentes a estados psicológicos de pacientes. Tendo também uma base lexical, que associa termos a estados emocionais, pretende-se desenvolver uma plataforma que permita ao utilizador uma pesquisa facilitada de documentos ou dados relativos aos pacientes, de forma a perceber o seu estado emocional.

1 Introdução

No âmbito da unidade curricular de *Descrição, Armazenamento e Pesquisa de Informação*, pretende-se permitir a pesquisa customizada e facilitada de documentos, a partir de vários parâmetros definidos pelo utilizador final. Esta pesquisa reflete sobre um *dataset* constituído por dados textuais relativos a consultas de vários pacientes, no âmbito da saúde mental, incluindo também classificações numéricas sobre várias métricas relativas ao bem-estar dos pacientes. A utilização de técnicas de *information retrieval* sobre estes dados permite também fazer uma extração e análise do estado psicológico do paciente, recolhendo informação não evidente que poderá permitir a análise mais completa do mesmo.

Este projeto surge como uma oportunidade de estudar e aproveitar dados provenientes de dois projetos diferentes mas relacionados entre si, o projeto *E-COMPARED* e o projeto *STOP DEPRESSION*. 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 o projeto da unidade curricular.

Este projeto pode também ser classificado 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.

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 espaço 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 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 foi traduzida a partir de um projeto francês designado *EMOTAIX*[5], e com-

posta por uma base lexical portuguesa designada de *PROLEX*. 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.

Existem diversas anotações no texto que devem não correspondem a elementos puramente textuais tais como:

- *C:* - intervenção por parte do paciente
- *P:* - intervenção por parte do terapeuta
- *(+: 00:05:00)*, *(+: 00:30:00)*, .. - tempo de gravação
- *(suspiro)*, *(risos)*, *(barulho externo)*, .. - ação
- *[(incompensível)]* - falha de transcrição

Existem também elementos simbólicos para os quais não é ainda conhecido o significado no contexto deste projeto:

- */*, *//*, *///*
- *:*
- +
- - - -, - -

Segundo esta estrutura a transcrição segue a seguinte forma, considerando o exemplo:

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.

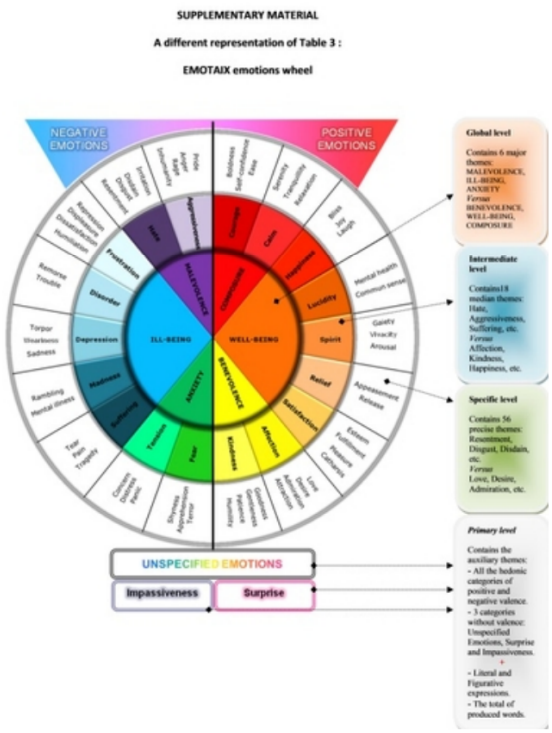

Figure 3: *Níveis da base lexical*

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

Figure 4: *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.

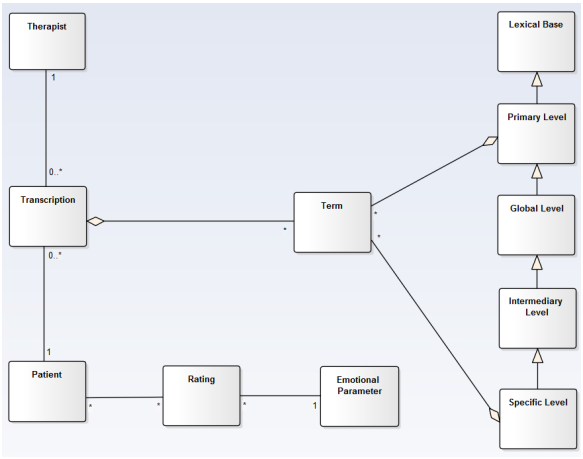

Figure 5: *Modelo conceptual*

4 *Information retrieval tasks*

A aplicação a desenvolver será uma ferramenta de pesquisa sobre os *datasets* referidos, retornando documentos mais adequados às preferências do utilizador. As pesquisas efetuadas podem ser aplicadas aos documentos, mas também a pacientes relacionados a esses documentos. Para além do retorno desta informação, pretende-se fazer uma pequena análise e aglomeração dos dados, de forma mais relevante para a pesquisa. Por exemplo, ao pesquisar por um dado termo em documentos, ao lado de cada documento retornado, poderá surgir um gráfico com a evolução do nível primário (emoções positivas/negativas) ao longo dos documentos associados a esse paciente, para ter uma percepção generalizada da evolução do paciente, e possível avaliação atual do seu estado psicológico, adaptando assim ao tema de *sentiment analysis*. Alguns exemplos de *information retrieval* interessantes pensados foram os seguintes:

- pesquisa por termos, ordenado pela ordem de aparência no documento ou documentos de um dado paciente. Pode ser selecionado se se pretende os documentos ou os pacientes cujo termo aparece mais ou menos vezes.

- pesquisa por níveis da base lexical ordenadamente.
  Exemplo: pacientes/documentos com mais emoções negativas ou positivas. Pode ser efetuada a pesquisa em outros níveis da base lexical, não sendo necessariamente o nível primário. Também esta pesquisa pode ser filtrada por paciente (perceber qual o paciente em melhor/pior estado psicológico), ou ser efetuada ao nível do documento.
- pesquisa por níveis da base lexical com um valor percentual inferior/superior a um dado limite.
  Exemplo: pacientes/documentos com emoções negativas com um valor superior a 50%.
- pesquisa pelos ratings ordenadamente.
  Exemplo: pacientes com rating de *mood* mais elevado (média dos valores), com maior evolução ao longo do tempo (média dos incrementos entre os valores). Esta pesquisa retorna pacientes, pois os *ratings* estão associados a estes, e não aos documentos.

5 Conclusões

Este trabalho poderá vir a ter resultados bastantes interessantes, facilitando a pesquisa para o utilizador final, sendo possível associar alguns dados retornados ao estado psicológico dos pacientes.

O carácter subjetivo das transcrições e respetiva dificuldade de análise de texto poderão ser limitações da eficiência e dos resultados obtidos, assim como limitações das ferramentas a ser utilizadas.

Considere-se o seguinte exemplo de excerto de uma transcrição *"C: Estou triste"*, comparativamente com a transcrição *"P: Está triste? / C: Sim."* poderá não ser de fácil associação pelas ferramentas utilizadas. Ignorar por completo as falas do terapeuta não será uma boa opção, mas conseguir associá-las ao estado do paciente poderá também não ser tarefa trivial.

References

[1] 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.

[2] 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.

[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*, páginas 1—-135, 2008.

[5] 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*, 2009.