Social Media Text Processing and Semantic Analysis for Smart Cities

João Filipe Figueiredo Pereira

Projecto de Dissertação de Mestrado desenvolvido com a orientação do *Prof. Rosaldo Rossetti* e *Pedro Saleiro* no *LIACC*

1. Motivação

Devido à ascensão das Redes Sociais, as pessoas obtêm e partilham informação quase que instantaneamente 24/7. Muitas áreas de investigação tentaram extrair informações importantes destes grandes volumes de conteúdo, gerado por utilizadores, e livremente disponíveis. As áreas de investigação de sistemas inteligentes de transportes e de cidades inteligentes (*smart cities*) não são excepção. Contudo, extrair conhecimento acionável e significativo de conteúdo gerado por utilizadores exige um esforço complexo. Primeiro, cada serviço de social media possui as suas próprias especificidades e restrições para o método de recolha dos dados; em segundo lugar, o volume de mensagens produzidas pode ser esmagador para o processamento automático e prospeção; e por último, não menos importante, os textos das redes sociais são, geralmente, curtos, informais, com muitas abreviações, jargões, gírias e expressões idiomáticas.

2. Descrição do Problema

Extrair conhecimento de dados do Twitter é um processo trabalhoso e demorado devido às restrições e dificuldades do conteúdo das mensagens. A informalidade, a existência de gírias, abreviaturas, jargões, e o curto comprimento das mensagens são alguns dos problemas emergidos durante a análise deste tipo de dados. Recolher *tweets* de forma automática e, ao mesmo tempo, extrair informação valiosa para cidades inteligentes e para a área de transportes torna a tarefa ainda mais complexa. A absência de *gold-standard datasets* é o problema mais perturbador, uma vez que não somos capazes de comparar as análises realizadas para as áreas de estudo acima mencionados. O problema em foco nesta dissertação é encontrar uma forma de demonstrar análises de texto das redes sociais sobre cidades/regiões/países que possam ser valiosas para entidades, governos ou até cidadãos comuns em processos de tomada de decisão, como, por exemplo, qual a cidade com nível de segurança mais elevado, ou qual o melhor meio de transporte que um indivíduo pode escolher para viajar em uma cidade.

3. Objectivos

Foram estabelecidos nove objectivos diferentes nesta dissertação para atingirmos o nosso alvo principal:

- Recolha contínua de *tweets* georreferenciados de múltiplas *bounding boxes* em paralelo e de acordo com os limites de utilização da *API* do *Twitter*
- Abordar as inconsistências e as filtragens de *tweets* com ruído da *API Twitter Geo*
- Implementar métodos padrão de pré-processamento de texto para mensagens das redes sociais.
- Análise de conteúdo recorrendo a modelação de tópicos e caracterização comparativa entre diferentes *bounding boxes* (por exemplo, cidades)
- Utilização de aprendizagem supervisionada na classificação de *tweets* relacionados com viagens e transportes
- Treino de *word embeddings* a partir de *tweets* georreferenciados
- Estudar o impacto de *word embeddings* em tarefas de classificação de *tweets* relacionados com viagens e transportes
- Criação de dados *gold-standard* para aprendizagem supervisionada relacionada com viagens e transportes
- Agregação e visualização de resultados

4. Proposta de Solução

Nesta secção é descrito o problema a ser abordado nesta dissertação bem como a arquitectura e a implementação da framework por nós proposta para resolução do problema apresentado, assim como os seus principais módulos constituintes.

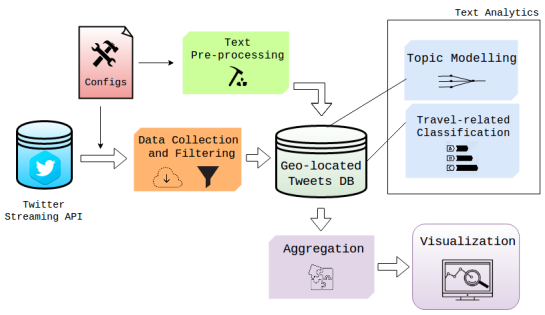

Fig. 1 – Visão geral da arquitectura da *framework*

4.1. Recolha e Filtragem de Dados

O módulo de recolha de dados foi desenvolvido recorrendo a uma biblioteca *open-source* de Python, conhecida como *Tweepy*, que permite o acesso às *APIs* do Twitter. Nós explorámos a *Twitter Streaming API* usando heurísticas de localização, que permitem a recolha de *tweets* georreferenciados através de uma sobreposição com *bounding-boxes*.

4.2. Pre-processamento de Texto

Nós aplicamos um grupo considerável de operaões de pré-processamento de texto às mensagens tais como *lowercasing*, *lemmatization*, *tokenization*, transformação de characters repetidos, filtragem de metadados, símbolos numéricos no texto e também remoção de palavras curtas e comuns.

4.3. Modelação de Tópicos

Redes sociais, mais especificamente, micro-blogs são serviços onde as pessoas publicam e partilham as suas opiniões e por esta razão são vistos como uma rica fonte de dados a explorar. De forma a explorar informação deste tipo de dados, nós implementámos na nossa *framework* um módulo generativo utilizando técnicas de modelação de tópicos. Nós escolhemos o modelo *Latent Dirichlet Allocation* [1] para a implementação do módulo e realizamos testes em duas cidades Brazileiras, Rio de Janeiro e São Paulo, utilizando um total de 9.5M de tweets georreferenciados. No final, foram descobertos 29 tópicos latentes (Figure 2) diferentes que caracterizam as duas cidades, sendo que 2 deles são únicos para cada cidade.

Fig. 2 – Tópicos latentes para as cidades do Rio de Janeiro e São Paulo

4.4. Classificação de Tweets sobre Transportes

Tentamos extrair e caracterizar tweets relacionados com viagens e transportes de grandes conjuntos de dados, a fim de estudar as distribuições geográficas e temporais deste conteúdo específico. As entidades de transporte poderão tirar proveito desse tipo de informação, uma vez que podem estudar os padrões de mobilidade humana, bem como as opiniões dos cidadãos sobre determinados serviços de transporte. Realizamos experiências no Rio de Janeiro, São Paulo e Nova Iorque para discriminar tweets relacionados a viagens e transportes. Devido ao volume de dados recolhidos para cada cenário, foi necessário identificar automaticamente os tweets cujo conteúdo de alguma maneira sugere estar relacionado ao domínio de transportes. As abordagens convencionais exigem que especifiquemos palavras-chave relacionadas a viagens e transportes para classificar *tweets*. Por outro lado, *word embeddings* [2] é uma forma de representação de texto que tenta capturar relações sintáticas e semânticas das palavras. O resultado é uma representação mais coesa onde palavras semelhantes são representadas por vectores semelhantes. Por exemplo, *"taxi"/"uber"*,*"busão/ônibus"*, *"ir para o trabalho"/"ir para a escola"* são casos em que os vetores produzidos são semelhantes. Foram estabelecidos diferentes grupos de *features* para treinar os nossos classificares de texto, nomeadamente *bag-of-words*, *bag-of-embeddings* e ambos, combinando horizontalmente as suas matrizes de representação de texto.

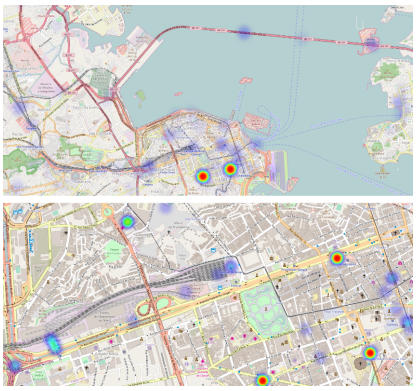

Fig. 3 – Aglomerações de *tweets* classificados como relacionados com transportes na cidade do Rio de Janeiro

5. Conclusão

Nesta dissertação tentamos enfrentar alguns dos desafios acima mencionados com o objetivo de extrair conhecimento de múltiplos fluxos de dados das redes sociais que possam ser úteis no contexto de sistemas inteligentes de transporte e cidades inteligentes. Criamos e desenvolvemos uma estrutura para recolha, processamento e exploração de Tweets geo-localizados. Mais especificamente, a ferramenta final fornece funcionalidades para a recolha paralela de tweets georreferenciados de várias *bounding-boxes* pré-definidas (cidades ou regiões), incluindo filtragem de *tweets* vazios, pré-processamento de texto para a língua portuguesa e inglesa, modelagem de tópicos e classificadores de texto relacionados com o domínio de transportes, bem como, agregação e visualização de dados.

Referências

[1] David M Blei, Andrew Y Ng, e Michael I Jordan. Latent dirichlet allocation. volume 3, páginas 993–1022, 2003.

[2] Tomas Mikolov, Wen-tau Yih, e Geoffrey Zweig. Linguistic regularities in continuous space word representations. Em *Hlt-naacl*, volume 13, 2013.