# Building a Question-Answering Corpus using Social Media and News Articles

Paulo Cavalin, Flavio Figueiredo, Maíra de Bayser,
Luis Moyano, Heloisa Candello, Ana Appel, and Renan Souza

IBM Research - Brazil

**Abstract.** Is it possible to develop a reliable QA-Corpus using social media data? What are the challenges faced when attempting such a task? In this paper, we discuss these questions and present our findings when developing a QA-Corpus on the topic of Brazilian finance. In order to populate our corpus, we relied on opinions from experts on Brazilian finance that are active on the Twitter application. From these experts, we extracted information from news websites that are used as answers in the corpus. Moreover, to effectively provide rankings of answers to questions, we employ novel word vector based similarity measures between short sentences (that accounts for both questions and Tweets). We validated our methods on a recently released dataset of similarity between short Portuguese sentences. Finally, we also discuss the effectiveness of our approach when used to rank answers to questions from real users.

**Keywords:** Question and Answer, Social Media, Finance

## 1 Introduction

The avalability of corpora to drive and sustain Question-Answering (QA) systems [8] is of fundamental importance. Such corpora are generally obtained from various sources, normally large collections of text, such as online news [1], or Wikipedia [6]. Some authors have put forward the advantages of using social media in the construction of certain types of corpora, e.g., in [3], the authors propose building comparable corpora from social networks, in particular, Twitter. In the same line, the authors of [5] propose using Twitter as an alternative in short sentences settings. Finally, authors in [7] address QA in social media mainly as a characterization effort on the type and frequency of questions and answers that may be found in Twitter, even though they don't address the construction of a corpus nor the targeting of any specific domain.

Motivated by the above setting, we study the potential of using social media data to create a QA corpus for specific fields. We refer to the corpus simply as QA-Corpus. In details, we study the viability of how can social media information be explored to create a QA-Corpus on the topic of Brazilian finance. To our knowledge, this is the first work to study the possibility of building a domain-specific corpus from social media. We believe that social media data can provide a proxy to field experts who can provide timely, possibly spam free

(using the right techniques), easy to understand, reliable information [3, 5, 11]. As previous work have showed, this information can be acquired by the careful choice of domain experts to follow in magazines, newspapers, or in a social media application such as Twitter [11]. In this sense, using social media we can bypass or tackle a major bottleneck on the creation of a QA-Corpus, of the gathering of candidate and reliable answers to possible user questions.

We combine the use of social media data together with novel, word vector based [4], short-sentence similarity measures to create a QA-Corpus that can answer user question in free text form. Using the method of [4], we match questions on Twitter, containing a URLs pointing out to a news article with the supposed answer, to real user questions.

We can summarize the major contribution of this work as the creation and evaluation of a method that automatically creates a QA-Corpus using social media data, based only on a seed set of Twitter ids. Using a novel similarity measure, this corpus can be used to provide answer to questions from real world users. We evaluate our steps on a recently released dataset of similarity between short Portuguese sentences. More importantly, we also show how QA-Corpus provides significant answers to questions provided from a user study performed by us. Even though we deal with texts in Portuguese and for the financial domain as a case study, the system here described can be easily applied to other languages and domains.

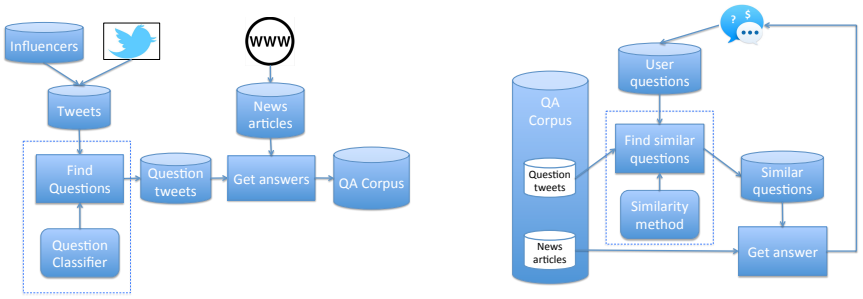
## 2 QA-Corpus Creation

In this section we describe the steps we took in building our QA-Corpus using social media data. Starting with Figure 1a, we present an overview of the main workflow of our QA-Corpus creation method. The main input is a list of social media accounts of experts of a given subject.

Given these users, the system collects all posts they submit on the social media service, for instance Twitter, and save them into a database, namely *Tweets DB*. Then, the module *Find Questions*, with the aid of the *Question Classifier*, finds all tweets that contain a question, and saves them into *Question tweets*. Finally, the *Get answers* module extracts the URL[1] of the news article linked in the text, and saves both the question tweet and its corresponding news article that answer the question into the QA-Corpus.

Our first approach to develop the question classifier was to train a supervised classifier to identify questions from tweets. However, as it has been discussed by previous work, simple heuristics can achieve over 90% accuracy [7] given the short nature of microblogging texts. Therefore, we considered a text as a question if it contains the question mark symbol ('?'). In addition, in order to populate our corpus with answers, we consider that any URL that follows a question on the tweet text is a candidate answer to that question.

[1] Uniform Resource Locator




a b

Fig. 2: In a), the main workflow of the QA-Corpus creation method making use of social media posts and news articles from the web. And in b), an illustration of the use of the QA-Corpus in a question answering system.

### 2.1 Social Media Data Extraction

To generate a corpus of social posts that is used as input for the QA-Corpus creation method, we have manually-selected a list with 104 Twitter users that are considered experts in Brazilian finance. This list includes journalists, bloggers, and professors. By means of the Twitter REST API, up to 30 November 2015, 184,001 posts could be collected. After applying the question classifier on these posts, the resulting QA-Corpus contained a total of 18,491 pairs of questions and answers, which corresponds to 10% of the total of posts.

## 3 Using QA-Corpus for a QA system

In this section, we present the results of using the QA-Corpus. described in Section 2.

For doing so, we rely on real user questions extracted from our alpha version QA financial application. In details, we performed a user study with 7 users, following the Wizard of Oz protocol [2], and extracted 124 questions on the topics of two savings investments available in most Brazilian banks: Savings Account and CDB (Bank Deposit Certificate) investments.

Moreover, we employ a word vector based similarity measure to find answers to questions on QA-Corpus. We validate the similarity measure on a dataset of similarity scores between pairs of Portuguese sentences. Finally, we also show how this method provides accurate answers to the real user questions.

### 3.1 QA System

Figure 1b depicts our proposed QA system based on the use of the QA-Corpus created in Section 2. The method works as follows. Given a question from the

*User questions* database, the module *Find similar questions*, with the aid of *Similarity method*, looks for the question tweets which are the most similar ones to the user question, and saves them into the *Similar questions* set. Next, the *Get answer* module gets corresponding news article that are related to the similar questions found in the *Question tweets* dataset. These news articles are then retrieved.

It is worth mentioning that all texts are pre-processed in the following way. First, the text is case-normalized, then it is tokenized. Next, all hashtags (tokens starting with an #) and URLs are removed.

### 3.2 Similarity Method

Our similarity method is based on the approach described in [4], making use of word vector representations, which is currently a well-known deep learning approach to extract the semantic meaning of the words [9, 10]. In this work, though, a regression model has been trained instead of a binary classifier, so that continous values of similarity can be used to rank the most similar sentences. In this case, texts with higher values are considered as more similar.

The regression model has been trained in the ASSIN similarity dataset, which has been released as part of the PROPOR Semantic Similarity and Textual Inference[2].

We made use of the set of 3,000 pairs of sentences in Brazilian portuguese to train a Support Vector Regression (SVR) model by considering a 30-dimensional feature set defined with both domain-independent data, from word vectors created with most recent dump of the Wikipedia in Portuguese[3], plus domain-specific data coming from all of the 184,001 originally crawled tweets and 64,646 news articles.

Based on a 70/30 division of the set, the SVR achieved Person correlation of 0.61, and mean squared error of 0.47. By considering a 3.5 threshold to convert the set into a binary classification problem, an SVM trained with the same configuration reached an F-score of 0.65.

### 3.3 Results

In Figure 3, we present the histogram according to the number of similar questions found for the 124 user questions. We can observe that, for the largest portion of questions, i.e. 47 questions or 38%, 0 to 4 similar questions are found in the dataset. From these, the system has not been able to find any similar question for 19 questions (40%). On the other tail, we observe that 39 questions had more than 20 similar questions, being in some cases very large sets ranging from 105 to 214 similar questions.

[2] ASSIN: Avaliação de Similaridade Semântica e Inferência Textual - http://propor2016.di.fc.ul.pt/?page$_id = 381$

[3] Dump of 12 December 2015

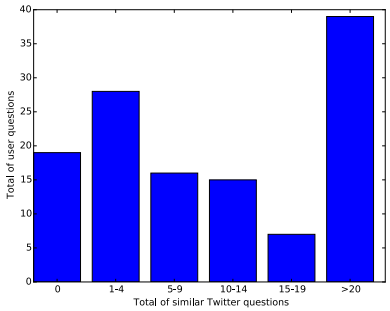



Fig. 3: Histogram of the number of similar questions found for the user questions.

In order to better understand the results, we have also conducted a qualitative analysis regarding the most similar questions that were found for each user question. We observe that the QA system can find pretty good matches when the question is very clear and direct, such as *What is CDB?*. Even when the question is not very clear, for instance *o rendimento em poupança é melhor*, the system has been able to point out some similar questions that might link to an answer. In some cases such as *qual a diferença entre poupança e cdb?* and *Então em curto prazo a poupança é mais rentável?* show that, even though the similarity method has captured the main meaning behind the question, lack of data probably contributed to not bringing any accurate similar question. Finally, no similar question was found for *investimento na poupana é seguro, mas existem outras opções que também são de baixo risco, mas com rentabilidade melhor.*, and the reason is not straightforward. This may have happend because: a) the question is not very clear; b) the text is too long; c) lack of data; or d) it is not a question but only a opinion.

## 4 Conclusions and Future Work

In this paper we presented a methodology to automatically create a QA-Corpus on the topic of Brazilian finance. The QA-Corpus is built through the use of social media data. We employ novel deep-learning based similarity measures to match questions from users and rank candidate answers. We validate our method on a novel dataset, as well as present a qualitative discussion of how our QA-Corpus can benefit real users for a financial advisor application.

As we have shown, with simple heuristics and state-of-the-art similarity measures, we can create the QA-Corpus. Nevertheless, one direction for future work is to improve the selection of social media posts to be included in QA-Corpus. Currently, we make use of every post in the form of a question. The aim is to increase the possibility of finding better answers to more specific user questions. Also, investigating other similarity measures for comparisons is a promising task for further study.

## References

1. Dolan, B., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th international conference on Computational Linguistics. p. 350. Association for Computational Linguistics (2004)
2. Dow, S.P., Mehta, M., MacIntyre, B., Mateas, M.: Eliza meets the wizard-of-oz: Blending machine and human control of embodied characters. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 547–556. ACM (2010)
3. Hajjem, M., Trabelsi, M., Latiri, C.: Building comparable corpora from social networks. In: BUCC, 7th Workshop on Building and Using Comparable Corpora, LREC, Reykjavik, Iceland (2013)
4. Kenter, T., de Rijke, M.: Short text similarity with word embeddings. In: CIKM 2015: 24th ACM Conference on Information and Knowledge Management. ACM (October 2015)
5. Ljubešic, N., Fišer, D., Erjavec, T.: Tweet-cat: a tool for building twitter corpora of smaller languages. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14), Reykjavik, Iceland. European Language Resources Association (ELRA) (2014)
6. Nothman, J., Murphy, T., Curran, J.R.: Analysing wikipedia and gold-standard corpora for ner training. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. pp. 612–620. Association for Computational Linguistics (2009)
7. Paul, S., Hong, L., Chi, E.: Is twitter a good place for asking questions? a characterization study. In: International AAAI Conference on Web and Social Media (2011)
8. Singh, V., Dwivedi, S.K.: Question answering: A survey of research, techniques and issues. International Journal of Information Retrieval Research (IJIRR) 4(3), 14–33 (2014)
9. Socher, R., Chen, D., Manning, C.D., Ng, A.: Reasoning with neural tensor networks for knowledge base completion. In: Advances in Neural Information Processing Systems. pp. 926–934 (2013)
10. Socher, R., Huang, E.H., Pennin, J., Manning, C.D., Ng, A.Y.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In: Advances in Neural Information Processing Systems. pp. 801–809 (2011)
11. Zafar, M.B., Bhattacharya, P., Ganguly, N., Gummadi, K.P., Ghosh, S.: Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream. ACM Transactions on the Web (TWEB) 9(3), 12 (2015)