DUTH at SemEval-2019 Task 8: Part-Of-Speech Features for Question Classification

Anastasios Bairaktaris **Symeon Symeonidis** **Avi Arampatzis**

Database and Information Retrieval research unit,
Department of Electrical and Computer Engineering,
Democritus University of Thrace, Xanthi 67100, Greece

{anasbair1,ssymeoni,avi}@ee.duth.gr

Abstract

This report describes the methods employed by the Democritus University of Thrace (DUTH) team for participating in SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums. Our team dealt only with Subtask A: Question Classification. Our approach was based on shallow natural language processing (NLP) preprocessing techniques to reduce noise in data, feature selection methods, and supervised machine learning algorithms such as NearestCentroid, Perceptron, and LinearSVC. To determine the essential features, we were aided by exploratory data analysis and visualizations. In order to improve classification accuracy, we developed a customized list of stopwords, retaining some opinion- and fact-denoting common function words which would have been removed by standard stoplisting. Furthermore, we examined the usefulness of part-of-speech (POS) categories for the task; by trying to remove nouns and adjectives, we found some evidence that verbs are a valuable POS category for the opinion-oriented question class.

1 Introduction

The significance of Community Question Answering (CQA) forums has risen in the past years. Such forums represent a modern need for information that comes with the abundance of online sources and the needs of millions of people for answers. Popular forums like StackOverflow, Yahoo! Answers, and Answers.com provide platforms for general or specific questions in a wide range of topics by users' and also a community-based model for user interaction.

The large numbers of questions and answers located in these forums generate many opportunities for information retrieval and data mining applications, such as query-intent detection, opinion mining, fake news classification, etc. (Tsur et al., 2016; Jo et al., 2018; Sethi, 2017). More advanced applications do not only aim at analyzing opinions but—by categorizing the feelings of the Q&As—they may be able to detect inappropriate content such as hate speech and act accordingly (Karadzhov et al., 2017; Baly et al., 2018).

The SemEval Task 8, Fact Checking in Community Forums, aims to determine whether the answers that are provided for a question in a forum are true or false. While answers to fact-oriented questions can be deemed true or false, opinion-oriented and socializing questions evoke answers for which a true/false categorization does not make much sense. As a result, determining the question type is a necessary first step. Consequently, the subtask A of SemEval Task 8 has the goal of classifying questions in three categories: opinion, factual, or socializing.

The rest of this report is structured as follows. Section 2 reviews some previous studies for CQA classification. Section 3 describes our system, while Section 4 presents experiments and results. Conclusions are summarized in Section 5.

2 Related Work

In recent years, plenty of research work examined the problem of classifying texts of CQA forums. Some related work which we found useful or inspiring are mentioned below.

Mihaylova et al. (2018) proposed a novel approach based on multi-faceted modeling of facts, which integrates knowledge from several complementary sources, such as the answer content (what is said and how), the author profile (who says it), the remainder of the community forum (where it is said), and external authoritative sources of information (external support).

Another study which provided us with helpful information about the importance of feature se-

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1155–1159
Minneapolis, Minnesota, USA, June 6–7, 2019.

lection on the development of a question classifier was by Huang et al. (2008). They demonstrated the importance of using the wh-word (what, which, when) in question classification. Such words are commonly disregarded and used in stopwords lists. Our approach is also trying to use features such as imperative verbs that indicate an opinion.

The SemEval-2015 Task 3, Answer Selection in Community Question Answering, targeted to classify comments in a thread as relevant, potentially useful, or bad, concerning the thread question (Nakov et al., 2015). This task encouraged solutions for the question classification problem that involved semantic or complex linguistic information.

Finally, (Mihaylova et al., 2016; Baldwin et al., 2016; Franco-Salvador et al., 2016) participated in subtasks A, B, and C at SemEval-2016 Task 3 that involved tasks for Question-Comment Similarity, Question-Question Similarity, and Question-External Comment Similarity. They proposed classification models and provided results that highlighted the importance of lexical and semantic features.

The aforementioned studies help to identify 'gaps' in this research topic and ways to attempt new and different approaches for question classification.

3 System Description

In this section, we give the details of our question classification model, applied pre-processing techniques, as well as some statistics and visualizations for the dataset of the task.

3.1 Dataset

The organizers provided the dataset in an XML format. The given training set consisted of 1,118 questions for Subtask A that were selected from the Qatar Living forum.

We used Python's Element Tree library to parse and isolate specific content from the XML. The interesting tags to select were RelQBody (the question) and RELQ_FACT_LABEL (labeled question by organizers).

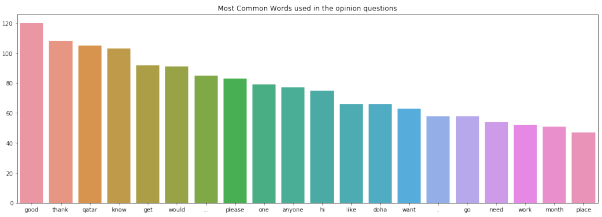
Before pre-processing, an exploratory data analysis gives us the opportunity to better understand the dataset. Because we will develop a multipurpose model that classifies not only the opinion but fact and socializing questions, it is helpful to understand in depth the character of the questions.

A way to understand the contents of the forum is to examine Table 1 where almost 50% of the questions are opinion oriented. Also, Figure 1 presents the most common words in opinion questions.

Label	Number of Questions
Opinion	563
Factual	311
Socializing	244

Table 1: Question types in the dataset

Figure 1: Most common words in opinion questions



3.2 Pre-processing

To reduce the noise of the text, based on the results of Symeonidis et al. (2018), we applied the following pre-processing:

- Remove Numbers
- Remove Punctuation
- Remove Symbols
- Lowercase
- Replace all URL addresses, normalizing them to 'URL'

Figure 2 shows the most frequent words on the dataset as a wordcloud.

The final steps of pre-processing are tokenization and stemming. A basic process in NLP is to identify tokens or those basic units which need not be decomposed in subsequent processing.

The entity word is one kind of token for NLP (Webster and Kit, 1992). Stemming is a process of reducing words to their stems or roots to reduce the vocabulary size and manage the case of data sparseness (Lin and He, 2009). For example, conjugated verbs such as 'goes', 'going', and 'gone' are stemmed to the term 'go'.

Figure 2: Wordcloud of frequent words

We used Python's SpaCy[1] library to tokenize the text and convert it to lemmas. This function also removes symbols (or punctuation) such as '[','...','-'.

Stopwords are frequent words that appear in the text, but they can have an impact on retrieval efficacy. The removal of stopwords also modifies the document length and subsequently affects the weighting process and efficiency during processing of the collection (Kwok, 1998).

For our task, we found out that the most commonly-used stopword lists contain words that can be helpful. For example, the word 'believe' is included in most stopword lists. While it is a ubiquitous word, it may also indicate an opinion; therefore, it can be useful for our purpose. In order to tackle this problem, we made a custom stopword list that only removes pronouns such as 'i','he','she', etc. NLTK's[2] list of English stopwords used as guideline and contained 127 words.

Although there is an abundance of stopwords lists that contain even more words we used a small one on purpose. We wanted to eliminate words from our dataset that would not bear any significance in our task. The next step, based on the vocabulary of the dataset, was to manually find words that could help us identify whether the question is opinion oriented, factual, or socializing. We excluded, from the NLTK's stopword list, words such as 'believe', 'think', 'mean', 'consider', and others. Our final revised stopword list consists of 50 words.

4 Experiments

This section summarizes our experiments in the context of SemEval 2019 Task 8 Subtask A. Beyond our officially submitted runs, we present some additional experiments that although they did not perform very well, there seems to be a promising room for improvement in the future.

4.1 Machine Learning Methods

For the training of our classifiers, we used Python's Scikit-Learn library (Pedregosa et al., 2011). We split the dataset into 749 training questions and 369 testing questions, i.e. a typical 2/3–1/3 split (ratio 2:1). After the split, the questions in the training set were shuffled for training. With the class `sklearn.pipeline`, we performed a sequence of different transformations and parameters.

Vectorizer: We compared three common vectorizers such as CountVectorizer, HashingVectorizer, and TfidfVectorizer. Finally, our selection was the TfidfVectorizer since it yielded the best results when it comes to accuracy. The TfidfVectorizer converts a collection of raw documents to a matrix of tf-idf weighted features.

Classifiers: We experimented with various classifiers, and decided to use the following three since they yielded the best accuracy results.

- **NearestCentroid**: Each class is represented by its centroid, with test samples classified to the class with the nearest centroid.
- **Perceptron**: It is a simple and efficient algorithm to fit linear models, and suitable for very large numbers of features.
- **LinearSVC**: An SVM algorithm which tries to find a set of hyperplanes that separate space into areas representing the classes. The hyperplanes are chosen in a way to maximize the distance from the nearest data point of each class.

4.2 Results

The official run that we submitted for the competition proved to be the most successful. In the following tables, we present the produced test results by using the three different classifiers, the TfidfVectorizer, and the custom stopword list. The results are shown in Tables 2, 3, and 4. We can observe that the most accurate classifier overall is the NearestCentroid.

We experimented further based on the hypothesis that opinion classification can be more effective by using only verbs. Until recently, most

[1] https://spacy.io
[2] https://www.nltk.org/

LinearSVC	accuracy	recall	f1-score
Factual	0.60	0.40	0.48
Opinion	0.62	0.82	0.70
Socializing	0.75	0.53	0.62
Total	0.64	0.64	0.62

Table 2: Test results with LinearSVC

NearestCentroid	accuracy	recall	f1-score
Factual	0.63	0.49	0.55
Opinion	0.71	0.76	0.73
Socializing	0.70	0.77	0.73
Total	0.68	0.69	0.68

Table 3: Test results with NearestCentroid

classification techniques have considered adjectives, adverbs, and nouns as features. The usefulness of part-of-speech categories in text classification was investigated as early as in (Arampatzis et al., 2000), where it was found that a traditional keyword-based indexing set can be reduced to retain only its nouns and adjectives without hurting effectiveness, even slightly improving it. Nevertheless, the aforementioned work was on topic classification; later, Karamibekr and Ghorbani (2012) showed that verbs are vital in classifying opinion terms, particularly in social domains.

We conducted two experiments by removing either nouns or adjectives from our dataset to help our classifiers adjust mostly on verbs. We can observe, in Tables 5 and 6, that classifiers achieved a better accuracy score when it comes to opinion as opposed to fact and socializing questions. Nevertheless, by removing either nouns or adjectives, there is an overall drop in effectiveness in all classes. Thus, there is evidence that verbs are a useful part-of-speech category for opinion classification, but they are not sufficient by themselves.

Our official submission to the competition ranked our team to the 16th place from 22 teams.

Perceptron	accuracy	recall	f1-score
Factual	0.54	0.36	0.43
Opinion	0.62	0.79	0.70
Socializing	0.64	0.49	0.56
Total	0.60	0.61	0.59

Table 4: Test results with Perceptron

NearestCentroid	accuracy	recall	f1-score
Factual	0.52	0.42	0.46
Opinion	0.68	0.66	0.67
Socializing	0.56	0.72	0.63
Total	0.61	0.61	0.60

Table 5: Test results without Nouns

NearestCentroid	accuracy	recall	f1-score
Factual	0.51	0.40	0.45
Opinion	0.66	0.65	0.66
Socializing	0.53	0.69	0.60
Total	0.59	0.59	0.59

Table 6: Test results without Adjectives

The results of our model are shown in Table 7.

Accuracy	F1	AverageRecall
0.71	0.56	0.60

Table 7: Official Results - Use of NearestCentroid

5 Conclusions

We presented a supervised learning model for classifying questions from online Q&A forums in three categories: factual, opinion, and socializing. We used standard pre-processing techniques, and made a custom stopword list to tackle the specific task at hand. Using standard classification methods, we achieved satisfactory and promising results. We also tried to use verb-oriented feature sets for classification which although they provided mixed results it seems that they can be improved.

References

Avi Arampatzis, Th.P. van der Weide, C.H.A. Koster, and P. van Bommel. 2000. An evaluation of linguistically-motivated indexing schemes. In *Proceedings of the 22nd BCS-IRSG Colloquium on IR Research*, pages 34–45.

Timothy Baldwin, Huizhi Liang, Bahar Salehi, Doris Hoogeveen, Yitong Li, and Long Duong. 2016. Unimelb at semeval-2016 task 3: Identifying similar questions by combining a CNN with string similarity measures. In (Bethard et al., 2016), pages 851–856.

Ramy Baly, Mitra Mohtarami, James R. Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov.

2018. Integrating stance detection and fact checking in a unified corpus. *CoRR*, abs/1804.08012.

Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch, editors. 2016. *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*. The Association for Computer Linguistics.

Marc Franco-Salvador, Sudipta Kar, Thamar Solorio, and Paolo Rosso. 2016. UH-PRHLT at semeval-2016 task 3: Combining lexical and semantic-based features for community question answering. In (Bethard et al., 2016), pages 814–821.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 927–936. ACL.

Saehan Jo, Immanuel Trummer, Weicheng Yu, Daniel Liu, and Niyati Mehta. 2018. The factchecker: Verifying text summaries of relational data sets. *CoRR*, abs/1804.07686.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully automated fact checking using external sources. *CoRR*, abs/1710.00341.

Mostafa Karamibekr and Ali A. Ghorbani. 2012. Verb oriented sentiment classification. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2012, Macau, China, December 4-7, 2012*, pages 327–331. IEEE Computer Society.

K. L. Kwok. 1998. Book review: Information storage and retrieval by r. r. korfhage. *Inf. Process. Manage.*, 34(4):490–492.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 375–384. ACM.

Tsvetomila Mihaylova, Pepa Gencheva, Martin Boyanov, Ivana Yovcheva, Todor Mihaylov, Momchil Hardalov, Yasen Kiprov, Daniel Balchev, Ivan Koychev, Preslav Nakov, Ivelina Nikolova, and Galia Angelova. 2016. Super team at semeval-2016 task 3: Building a feature-rich system for community question answering. In (Bethard et al., 2016), pages 836–843.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James R. Glass. 2018. Fact checking in community forums. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5309–5316. AAAI Press.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 269–281. The Association for Computer Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Ricky J. Sethi. 2017. Crowdsourcing the verification of fake news and alternative facts. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT 2017, Prague, Czech Republic, July 4-7, 2017*, pages 315–316. ACM.

Symeon Symeonidis, Dimitrios Effrosynidis, and Avi Arampatzis. 2018. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Syst. Appl.*, 110:298–310.

Gilad Tsur, Yuval Pinter, Idan Szpektor, and David Carmel. 2016. Identifying web queries with question intent. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 783–793. ACM.

Jonathan J. Webster and Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, August 23-28, 1992*, pages 1106–1110.