Topic-Specific Stylistic Variations for Opinion Retrieval on Twitter

Anastasia Giachanou[1], Morgan Harvey[2], and Fabio Crestani[1]

[1] Faculty of Informatics, Università della Svizzera italiana (USI), Switzerland
[2] Department of Maths and Information Sciences, Northumbria University, UK
{anastasia.giachanou,fabio.crestani}@usi.ch
morgan.harvey@northumbria.ac.uk

Abstract. Twitter has emerged as a popular platform for sharing information and expressing opinions. Twitter opinion retrieval is recognized as a powerful tool for finding people's attitudes on different topics. However, the vast amount of data and the informal language of tweets make opinion retrieval on Twitter very challenging. In this paper, we propose to leverage topic-specific stylistic variations to retrieve tweets that are both relevant and opinionated about a particular topic. Experimental results show that integrating topic specific textual meta-communications, such as emoticons and emphatic lengthening in a ranking function can significantly improve opinion retrieval performance on Twitter.

Keywords: opinion retrieval, microblogs, stylistic variations

1 Introduction

Microblogs have emerged as a popular platform for sharing information and expressing opinion. Twitter attracts 284 million active users per month who post about 500 million messages every day[3]. Due to its increasing popularity, Twitter has emerged as a vast repository of information and opinion on various topics. However, all this opinionated information is hidden within a vast amount of data and it is therefore impossible for a person to look through all data and extract useful information.

Twitter opinion retrieval aims to identify tweets that are both relevant to a user's query and express opinion about it. Twitter opinion retrieval can be used as a tool to understand public opinion about a specific topic, which is helpful for a variety of applications. One typical example refers to enterprises that can capture the views of customers about their product or their competitors. This information can be then used to improve the quality of their services or products accordingly. In addition, it is possible for the government to understand the public view regarding different social issues and act promptly.

Retrieving tweets that are opinionated about a specific topic is a non-trivial task. One of the many reasons is the informal nature of the medium, which has

[3] See: https://about.twitter.com/company/

effected the emergence of new stylistic conventions such as emoticons, emphatic lengthening and slang terms widely used on Twitter. These informal stylistic conventions can, however, be a valuable source of information when retrieving tweets that express opinion towards a topic. The use of emoticons usually implies an opinion [7] and emphatic lengthening has been shown to be strongly associated with opinionatedness [2]. For the rest of the paper, we use the phrases *stylistic conventions* and *stylistic variations* interchangeably to denote the emerged textual conventions in Twitter such as the emoticons and the emphatic lengthening. The stylistic variations are only a subset of the writing style of users in Twitter which refers to a much wider manner that is used in writing [18].

The extent to which stylistic variations are used varies considerably among the different topics discussed on Twitter. That is, the number of the stylistic variations present in each tweet is dependent on its topic. For example, tweets about entertainment topics (i.e. movies, TV series) tend to use more stylistic variations than those that express opinion about social issues (i.e. immigration) or products (i.e. Google glass). This implies that stylistic variations do not have the same importance in revealing opinion across different topics.

Here we propose a Twitter opinion retrieval model which uses information about the topics of tweets to retrieve those that are relevant and contain opinion about a user's query. The proposed model calculates opinionatedness by combining information from the tweet's terms and the topic-specific stylistic variations that are extensively used in Twitter. We compare several combinations of stylistic variations, including emoticons, emphatic lengthening, exclamation marks and opinionated hashtags, and evaluate the proposed model on the opinion retrieval dataset proposed by Luo et al. [9]. Results show that stylistic variations are topic-specific and that incorporating them in the ranking function significantly improves the performance of opinion retrieval on Twitter.

2 Related Work

With the rapid growth of social media platforms, sentiment analysis and opinion retrieval has attracted much attention in the research community. Early research focused on classifying documents as expressing either a positive or a negative opinion [14, 13]. A comprehensive review of opinion retrieval and sentiment analysis can be found in a survey by Pang and Lee [14].

The increasing popularity of Twitter has recently stirred up research in the field of Twitter sentiment analysis. One of the first studies was carried out by Go et al. [3] treated the problem as one of binary classification, classifying tweets as either positive or negative. Due to the difficulty of manually tagging the sentiment of the tweets, they employed distant supervision to train a supervised machine learning classifier. The authors used a technique devised by Read [17] to collect the data, according to which emoticons can be used to differentiate the negative and positive tweets. They compared Naive Bayes (NB), Maximum Entropy (MaxEnt) and Support Vector Machines (SVM), among which SVM with unigrams achieved the best result. Following Go et al. [3], Pak and Paroubek [11] used emoticons to label training data from which they built a multinomial Naïve Bayes classifier which used N-gram and POS-tags as features.

Due to the informal language used on Twitter, which frequently contains unique stylistic features, a number of researchers explored features such as emoticons, abbreviations and emphatic lengthening, studying their impact on sentiment analysis. Brody and Diakopoulos [2] showed that the lengthening of words (e.g., cooool) in microblogs is strongly associated with subjectivity and sentiment. Kouloumpis et al. [7] showed that Twitter-specific features such as the presence or absence of abbreviations and emoticons improve sentiment analysis performance. None of these approaches considered, however, the possibility that stylistic features may depend on the topic of the tweet.

Topic-dependent approaches have been considered by researchers in relation to terms. Jiang et al. [6] used manually-defined rules to detect the syntactic patterns that showed if a term was related to a specific object. They employed a binary SVM to apply subjectivity and polarity classification and utilised microblog-specific features to create a graph which reflects the similarities of tweets. Canneyt et al. [19] introduced a topic-specific classifier to effectively detect the tweets that express negative sentiment whereas Wang et al. [20] leveraged the co-occurrence of hashtags to detect their sentiment polarity.

Twitter opinion retrieval was first considered by Luo et al. [9] who proposed a learning-to-rank algorithm for ranking tweets based on their relevance and opinionatedness towards a topic. They used SVM$^{\text{Rank}}$ to compare different social and opinionatedness features and showed they can improve the performance of Twitter opinion retrieval. However, this improvement is over relevance baselines (BM25 and VSM retrieval models) and not over an opinion baseline. Our work is different as we propose to incorporate topic-specific stylistic variations into a ranking function to generate an opinion score for a tweet. To the best of our knowledge, there is no work exploring the importance of topic-specific stylistic variations for Twitter opinion retrieval. Another important difference is that we use both relevance and opinion baselines to compare the proposed topic-specific stylistic opinion retrieval method.

3 Topic Classification

Topic models aim to identify text patterns in document content. Standard topic models include Latent Dirichlet Allocation (LDA) [1] and Probabilistic Latent Semantic Indexing (pLSI) [4]. LDA, one of the most well known topic models, is a generative document model which uses a “bag of words” approach and treats each document as a vector of word counts. Each document is a mixture of topics and is represented by a multinomial distribution over those topics. More formally, each document d in the collection is associated with a multinomial distribution over K topics, denoted θ. Each topic z is associated with a multinomial distribution over words, denoted ϕ. Both θ and ϕ have Dirichlet prior with hyperparameters α and β respectively. For each word in a document d, a topic z is sampled from the multinomial distribution θ associated with the document and a word w from the multinomial distribution ϕ associated with topic z. This generative process is repeated N_d times, where N_d is the total number of words in the document d. LDA defines the following process for each document:

1. Choose $\theta_d \sim \text{Dir}(\alpha)$,
2. Choose $\phi_z \sim \text{Dir}(\beta)$,
3. For each of the N words w_n:
   (a) Pick a topic z_n from the multinomial distribution θ_d
   (b) Pick a word w_n from the multinomial distribution ϕ_z

Topic models have been applied in a wide range of areas including Twitter. Hong and Davison [5] conducted an empirical study to investigate the best way to train models for topic modeling on Twitter. They showed that topic models learned from aggregated messages of the same user may lead to superior performance in classification problems. Zhao et al. [22] proposed the Twitter-LDA model which considered the shortness of tweets to compare topics discussed in Twitter with those in traditional media. Ramage et al. [16] applied labeled-LDA in Twitter, a partially supervised learning model based on hashtags. Inspired by the popularity of LDA, Krestel et al. [8] proposed using LDA for tag recommendation. In this work, we use LDA [1] to determine the topics of tweets, which are then used to learn the importance of the stylistic variations for each topic.

4 Twitter Opinion Retrieval

Twitter opinion retrieval aims to develop an effective retrieval function which retrieves and ranks tweets accordingly to the likelihood that express an opinion about a particular topic. The proposed approaches for opinion retrieval usually follow a three step framework. In the first step, traditional IR methods are applied to rank documents by their relevance to the query. In the second, opinion scores are generated for the documents that were retrieved during the first step and, in the last step, a final ranking of the documents is produced based both on their relevance and opinionatedness towards the query.

In this section, we propose a new opinion retrieval model which leverages topic-specific stylistic variations of short informal texts such as tweets to calculate their opinionatedness. The proposed model calculates the opinionatedness of a document by combining two different opinion scores. The *term-based* component is based on the opinionatedness of the document's terms, whereas the *stylistic-based* component instead considers the stylistic variations present in the document.

Let $S_d(o)$ be the opinion score of a document (tweet) d based on its terms and $S_{ls,d}(o)$ be the opinion score of a document d based on the stylistic variations that d contains. Then the opinionatedness of the document d is the weighted sum of the two opinion score components and is calculated as follows:

$$S_{q,d}(o) = \lambda * S_d(o) + (1 - \lambda) * S_{ls,d}(o)$$

where $\lambda \in [0, 1]$.

Term-Based Opinion Score. The presence of opinionated terms in a document, and their probability of expressing opinion, is a popular approach to calculate the document's opinionatedness. A simple method is to calculate this score as the average opinion score over all terms in the document, thus:

$$S_d(o) = \sum_{t \in d} opinion(t) p(t|d) \tag{1}$$

where $p(t|d) = c(t,d)/|d|$ is the relative frequency of term t in document d and $opinion(t)$ shows the opinionatedness of the term.

Since this is one of the most widely used methods to calculate the opinionatedness of a document, we also use this method as one of our baselines.

Stylistic-Based Opinion Score. Our method incorporates several stylistic variations of tweets into a ranking function to rank tweets according to their opinionatedness. The stylistic-based component of our model calculates an opinion score using the stylistic variations that a document contains. Let l be a stylistic variation taken from the list $L = (l_1, ..., l_i, ..., l_{|l|})$ which includes all the possible stylistic variations that reveal opinions. We then calculate the stylistic-based component as follows:

$$S_{ls,d}(o) = \sum_{l \in LS} SVF(l,d) * IDF(l)$$

where LS is a subset of stylistic variations ($LS \subset L$), $SVF(l,d)$ represents the frequency of the stylistic variation l in the document d and $IDF(l)$ represents the importance of the variation l, that is if the stylistic variation is common across the documents or not. The inverse frequency IDF of the stylistic variation l controls the amount of opinion information that the specific variation holds.

We explore various ways of calculating the frequency SVF of the stylistic variations. These are the following:

$$SVF_{Bool}(l,d) = \begin{cases} 0, & \text{if } f(l,d) = 0 \\ 1, & \text{if } f(l,d) > 0 \end{cases}$$

$$SVF_{Freq}(l,d) = f(l,d)$$

$$SVF_{Log}(l,d) = 1 + \log f(l,d)$$

where $f(l,d)$ is the number of occurrences of variation l in document d.

To model the relative importance of each stylistic variation l across the documents we consider the following methods:

$$IDF_{Inv}(l) = \log \frac{N}{1 + n_l} \tag{2}$$

$$IDF_{Prob}(l) = \begin{cases} 0, & \text{if } N = n_l \\ \log \frac{N - n_l}{n_l}, & \text{if } N \neq n_l \end{cases} \tag{3}$$

where n_l can also be written as $|d \in D : l \in d|$ and denotes the number of documents that belong to collection D and contain the stylistic variation l. Thus, the importance of a given stylistic variation l depends on how frequently it is used in the collection D.

Topic Specific Stylistic-Based Opinion Score. The assumption made in the existing literature, that the stylistic variations are used with the same frequency across the documents of different topics, is not accurate. The informal stylistic variations are used with differing frequencies depending on the topic discussed. For example, tweets that are relevant to a TV series probably contain more stylistic variations than those that are relevant to a social issue, such as immigration. That means that the probability that stylistic variations imply opinion depends on the topic. In other words, if emoticons are extensively used in tweets about a specific topic, then their ability to imply opinion decreases.

Based on this assumption, we propose using topic-specific stylistic variations. To this end, we first apply topic modeling to determine the topic of a tweet and then we use this information to calculate the stylistic-based component of our approach, that is the opinionatedness of a tweet when it contains a specific stylistic variation. More formally, let $T = (T_1, ..., T_i, ..., T_{|T|})$ be the topics extracted after applying a topic model on the collection D, and $D_T = (d_1, ..., d_t)$ the documents that were assigned to the topic T_i. Then, the relative importance IDF of each stylistic variation l is calculated using the equations 2 and 3 with the difference that n_l denotes the number of documents that belong to collection D_T and contain the stylistic variation l. In other words, n_l is calculated as $|d \in D_T : l \in d|$, where D_T is a collection of documents that were assigned the same topic T_i.

Combining Relevance and Opinion Scores. To generate the final ranking of documents according to their relevance and opinionatedness, we combine the relevance score with the opinionatedness of the tweet:

$$S_{o,q}(d) = S_d(q) * S_{q,d}(o)$$

where $S_d(q)$ is the relevance score of d given topic t and $S_{q,d}(o)$ is the opinionatedness of d. $S_d(q)$ can be estimated using any existing IR model.

5 Experimental Setup

Dataset. To evaluate our methods we used the dataset proposed by Luo et al. [9], which is, to the best of our knowledge, the only dataset that has been used for Twitter opinion retrieval. The original collection contains 50 topics and 5000 judged tweets crawled in November 2011. We note that there is another dataset which can be used for opinion retrieval in Twitter. This dataset was created by Paltoglou and Buckley [12] who annotated part of the Microblog dataset provided by TREC with subjectivity annotations. However, as this dataset has not yet been used in any study, we would not be able to make direct comparisons of our methods and therefore only consider the first.

Experimental Settings. To create the index, we removed URLs, hashtag symbols (#) placed in front of some terms and character repetitions that appear more than twice in a row in a term. We indexed the collection with the Terrier IR system[4]. Our preprocessing also involves stop-word removal using the snowball stop word list[5] and stemming using the Porter stemmer [15].

To avoid overfitting the data we performed 5 fold cross-validation on the 50 queries. For each fold we used 40 queries for the training phase and 10 for testing. The training and test data was kept separate in all phases of our experiments. We perform our experiments under two different settings: *non topic-based* and *topic-based*. For the non topic-based settings, we apply the proposed method on the whole collection without considering the tweet's topic. For the topic-based settings we first apply LDA to detect the topics and then we apply the proposed method on tweets of the same topic. To estimate the LDA parameters we used a Gibbs sampler. Since the Gibbs sampler is a stochastic method, and therefore will produce different outputs by run, we report the mean performance of the methods based on ten runs.

Opinion Lexicon and Stylistic Variations. To identify the opinionated terms we use the AFINN Lexicon, as proposed by Nielsen [10]. AFINN contains more than 2000 words, each of which is assigned a valence from -5 to -1 for terms with a negative sentiment or from 1 to 5 for terms with a positive sentiment. We chose this lexicon as it contains affective words that are used in Twitter. We took the absolute values of the scores since we do not consider sentiment polarity in our study. We use $MinMax$ normalisation to convert the valence score of a term to opinion score. To avoid getting zero scores for terms with absolute score 1, we consider that the lexicon has also one term with no sentiment (assigned the score 0), so that 0 is the minimum score.

To calculate the stylistic-based component of our model, we identified, for each tweet, the number of emoticons, exclamation marks, terms under emphatic lengthening and opinionated hashtags as follows:

- *Emoticons*: Number of emoticons in a tweet. For the emoticons, we used the list provided on Wikipedia[6]. We consider all emoticons to be opinion-bearing. Therefore, we did not distinguish them by their subjectivity, sentiment or emotion they express.
- *Exclamation marks*: Number of exclamation marks in a tweet.
- *Emphatic lengthening*: Number of terms under emphatic lengthening in a tweet, that is terms that contain more than two repeated letters.
- *Opinionated hashtags*: Number of opinionated hashtags. As opinionated hashtags we considered any hashtag whose term is contained in the AFINN opinion lexicon. For example, the hashtag *#love* is considered an opinionated hashtag because the term *love* appears in the AFINN opinion lexicon.

[4] Available at: `http://terrier.org/`

[5] Available at: `http://snowball.tartarus.org/`

[6] See `http://en.wikipedia.org/wiki/List_of_emoticons`

Evaluation. We compare the proposed opinion retrieval method with two baselines. The first, *BM25*, is the method with the best performance in Twitter opinion retrieval according to the results presented in [9]. The *Relevance-Baseline* is based purely on topical relevance and does not consider opinion. As a second baseline, we use the term-based opinion score (equation 1). The *Opinion-Baseline* considers opinion and therefore it is a more appropriate baseline to compare our results with. To evaluate the methods, we report *Mean Average Precision* (MAP), which is the only metric reported in previous work [9] in Twitter opinion retrieval. To compare the different methods we used the Wilcoxon signed ranked matched pairs test with a confidence level of 0.05.

6 Results and Discussion

Topic Classification. In order to identify the topics discussed, we applied the LDA [1] topic model on the dataset proposed by Luo et al. [9]. For the analysis, we applied Gibbs sampling for the LDA model parameter estimation and inference as proposed in [21]. We considered each tweet to be a document. We tried a number of different values for the K parameter, which represents the number of topics, ranging from 1 to 200 with a step of 5. We set the number of iterations to 2000. The maximum log likelihood is obtained for 65 topics.

Table 1. Sample of topic descriptions when the number of topics is set to 65

Sample topics from Twitter
jennifer aniston lopez brad
steve jobs apple biography
disney world walt princess
music awards red carpet
biology chemistry science lab

Table 1 shows a list of five topics which were discovered in the collection of tweets when the number of topics was set to 65. We observe that LDA managed to group terms that are about the same topic together.

Twitter Opinion Retrieval. Table 2 presents the results of Twitter opinion retrieval when different stylistic variations are combined. Any of the approaches of calculating SVF and IDF presented in Section 4 can be used to evaluate the effectiveness of the different combinations. For the results displayed in Table 2 we applied SVF_{Log} and IDF_{Inv} under topic-based settings. We observe that all the three examined combinations (SVF$_{Log}$IDF$_{Inv}$-Emot-Excl, SVF$_{Log}$IDF$_{Inv}$-Emot-Excl-Emph, SVF$_{Log}$IDF$_{Inv}$-Emot-Excl-Emph-OpHash) perform significantly better than both the relevance and opinion baselines. Though there is no statistical difference between the different combinations of the stylistic variations, the best performance is achieved when we combined *emoticons*, *exclamation marks* and *emphatic lengthening*. This is a very interesting result that shows that integrating the most useful stylistic variations and the opinionatedness of the terms into a ranking function can be very effective for Twitter opinion retrieval.

Table 2. Performance results of the $\mathrm{SVF}_{Log}\mathrm{IDF}_{Inv}$ method under topic-based settings using different combinations of stylistic variations over the baselines. A star($*$) and dagger($\dagger$) indicate statistically significant improvement over the relevance and opinion baselines respectively.

	MAP
Relevance-Baseline	0.2835
Opinion-Baseline	0.3807$*$
$\mathrm{SVF}_{Log}\mathrm{IDF}_{Inv}$-Emot-Excl	0.4314$*$ $\dagger$
$\mathrm{SVF}_{Log}\mathrm{IDF}_{Inv}$-Emot-Excl-Emph	0.4413$*$ $\dagger$
$\mathrm{SVF}_{Log}\mathrm{IDF}_{Inv}$-Emot-Excl-Emph-OpHash	0.4344$*$ $\dagger$

Table 3 shows the performance of the proposed model on non topic-based and topic-based settings for Twitter opinion retrieval. We evaluate the effectiveness of different combinations of approaches in calculation of SVF and IDF. We observe that most of the approaches perform statistically better under the topic-based settings compared to the non topic-based settings. This is a very interesting result which shows that stylistic variations are indeed topic-specific and the amount of the opinion information they hold depends on the topic of the tweet. We also observe that there is no statistical difference between the different SVF and IDF approaches when they are compared under the same settings.

Table 3. Performance results of different SVF and IDF combinations, based on emoticons, exclamation marks and emphatic lengthening. A star($*$) indicates statistically significant improvement over the non topic-based settings for the same approach.

SVF - IDF	Non Topic-Based	Topic-Based
$\mathrm{SVF}_{Bool}\mathrm{IDF}_{Inv}$	0.4279	0.4419$*$
$\mathrm{SVF}_{Freq}\mathrm{IDF}_{Inv}$	0.4279	0.4398
$\mathrm{SVF}_{Log}\mathrm{IDF}_{Inv}$	0.4275	0.4413$*$
$\mathrm{SVF}_{Bool}\mathrm{IDF}_{Prob}$	0.4279	0.4427$*$
$\mathrm{SVF}_{Freq}\mathrm{IDF}_{Prob}$	0.4279	0.4421$*$
$\mathrm{SVF}_{Log}\mathrm{IDF}_{Prob}$	0.4275	0.4429$*$

In addition, we performed a per topic analysis to compare the model under topic-based against non topic-based settings. Table 4 shows the three topics that were helped or hurt the most using the $\mathrm{SVF}_{Log}\mathrm{IDF}_{Prob}$ model under the topic-based compared to the non topic-based settings. We observe that the topics that were helped are those that probably contain few informal stylistic variations as they are related to topics about products or politics. In future, we plan to do a thorough exploration to detect the possible reasons for the increase/decrease in the performance of the topics.

Table 4. Topics that are helped or hurt the most in the SVF$_{Log}$IDF$_{Prob}$ model under topic-based compared to non topic-based settings.

Helped		Hurt	
Title	ΔMAP	Title	ΔMAP
iran	0.1795	new start-ups	-0.1833
Lenovo	0.1185	iran nuclear	-0.0480
galaxy note	0.1017	big bang	-0.0319

Finally, we compare the performance of our proposed approach with the performance of the best run presented by Luo et al. [9] and report the comparison result in Table 5. We observe that our best runs outperform their best reported result (denoted BM25_Best). Finally, we should mention that their method uses SVM$^{\mathrm{Rank}}$ and their best run (BM25_Best) is trained using a number of social features (URL, Mention, Statuses, Followers) together with BM25 score, and Query-Depedent opinionatedness (Q_D) features.

Table 5. Results on Δ MAP for best runs over Opinion-Baseline

Run	Map	Δ MAP
Opinion-Baseline	0.3807	-
BM25_Best	0.4181	9.82%
SVF$_{Log}$IDF$_{Prob}$-Emot-Excl-Emph	0.4429	16.33%
SVF$_{Bool}$IDF$_{Prob}$-Emot-Excl-Emph	0.4427	16.28%

Post Analysis. In order to validate the results of our study we carry out a post analysis to measure the number of tweets that contain stylistic variations and the number of relevant tweets that contain stylistic variations per each topic. Let consider that T denotes the tweets that belong to a topic, STV represents the tweets that contain at least one stylistic variation and R the tweets that are relevant according to relevance judgments (*qrels*). Then, we can calculate the ratio of tweets that contain stylistic variations per topic as:

$$r_{stv}(T) = \frac{|d \in T \cap STV|}{|d \in T|}$$

and the ratio of relevant tweets that contain stylistic variations as:

$$r_{stv}(T,R) = \frac{|d \in T \cap STV \cap R|}{|d \in T \cap STV|}$$

Figure 1 shows the ratio of tweets that contain stylistic variations and of those that contain stylistic variations and are also relevant for each topic. The topics are sorted according to $r_{stv}(T)$ to better illustrate the relation of the two ratios. It can be seen that as the ratio of tweets with stylistic variations increases the ratio of relevant tweets with stylistic variations decreases. Therefore we can use topic specific stylistic variations to differentiate the relevant documents from non-relevant documents.

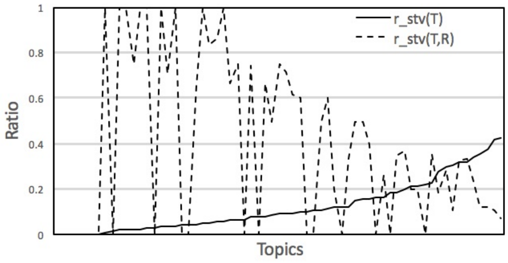

Fig. 1. Ratio of tweets with stylistic variations and of relevant tweets with stylistic variations per topic

7 Conclusions and Future Work

In this paper, we considered the problem of Twitter opinion retrieval. We proposed a topic-based method that uses topic-specific stylistic variations to address the problem of opinion retrieval in Twitter. We studied the effect of different approaches and of the different stylistic variations in the performance of Twitter opinion retrieval. The experimental results showed that stylistic variations are good indicators for identifying opinionated tweets and that opinion retrieval performance is improved when emoticons, exclamation marks and emphatic lengthening are taken into account. Additionally, we demonstrated that the importance of stylistic variations in indicating opinionatedness is indeed topic dependent as our topic model-based approaches significantly outperformed those that assumed importance to be uniform over topics.

In future, we plan to extend the topic-based opinion retrieval method by investigating the effect of assigning different importance weights to each stylistic variation. We also plan to evaluate the performance of our method on other datasets that consider opinion retrieval on short texts that share similar stylistic variations to tweets such as MySpace and YouTube comments.

Acknowledgments. This research was partially funded by the Swiss National Science Foundation (SNSF) under the project OpiTrack.

References

1. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning research 3, 993–1022 (2003)
2. Brody, S., Diakopoulos, N.: Cooooooollllllllll!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In: EMNLP'11. pp. 562–570 (2011)
3. Go, A., Bhayani, R., Huang, L.: Tech. rep., Standford (2009)
4. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: SIGIR'99. pp. 50–57 (1999)
5. Hong, L., Davison, B.D.: Empirical Study of Topic Modeling in Twitter. In: SIGKDD Workshop on SMA. pp. 80–88 (2010)
6. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: ACL'11. pp. 151–160. HLT '11 (2011)
7. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: ICWSM'11. pp. 538–541 (2011)
8. Krestel, R., Fankhauser, P., Nejdl, W.: Latent dirichlet allocation for tag recommendation. In: RecSys'09. pp. 61–68 (2009)
9. Luo, Z., Osborne, M., Wang, T.: An effective approach to tweets opinion retrieval. In: WWW'13. pp. 1–22 (2013)
10. Nielsen, F.Å.: A new ANEW: Evaluation of a word list for sentiment analysis of microblogs. In: ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. pp. 93–98 (2011)
11. Pak, A., Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: LREC'10. pp. 1320–1326 (2010)
12. Paltoglou, G., Buckley, K.: Subjectivity annotation of the microblog 2011 realtime adhoc relevance judgments. In: ECIR'13. pp. 344–355 (2013)
13. Paltoglou, G., Giachanou, A.: Opinion retrieval: Searching for opinions in social media. In: Paltoglou, G., Loizides, F., Hansen, P. (eds.) Professional Search in the Modern World, pp. 193–214. Springer International Publishing (2014)
14. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
15. Porter, M.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
16. Ramage, D., Dumais, S., Liebling, D.: Characterizing Microblogs with Topic Models. In: ICWSM'10. pp. 1–8 (2010)
17. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: ACL Student Research Workshop. pp. 43–48 (2005)
18. Strunk, W.: The elements of style. Penguin (2007)
19. Van Canneyt, S., Claeys, N., Dhoedt, B.: Topic-dependent sentiment classification on twitter. In: ECIR'15. pp. 441–446 (2015)
20. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach. In: CIKM '11. pp. 1031–1040 (2011)
21. Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. In: SIGKDD'09. pp. 937–946 (2009)
22. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., Li, X.: Comparing Twitter and Traditional Media using Topic Models. In: ECIR'11. pp. 338–349 (2011)