

LIA@CLEF 2018: Mining events opinion argumentation from raw unlabeled Twitter data using convolutional neural network^{*}

Richard Dufour¹, Mickaël Rouvier¹,
Alexandre Delorme², and Damien Malinas²

¹ LIA - University of Avignon (France)

{firstname.lastname}@univ-avignon.fr

² UMR 8562 Centre Norbert Elias - University of Avignon (France)

{firstname.lastname}@univ-avignon.fr

Abstract. Social networks on the Internet are becoming increasingly important in our society. In recent years, this type of media, through communication platforms such as Twitter, has brought new research issues due to the massive size of data exchanged and the important number of ever-increasing users. In this context, the CLEF 2018 Mining opinion argumentation task aims to retrieve, for a specific event (festival name or topic), the most diverse argumentative microblogs from a large collection of tweets about festivals in different languages. In this paper, we propose a four-step approach for extracting argumentative microblogs related to a specific query (or event) while no reference data is provided.

Keywords: Opinion detection · Microblogs · Unlabeled data · Convolutional neural network.

1 Introduction

Social networks on the Internet allow communities of users to exchange and share resources worldwide (ideas, opinions, data...) to an increasingly wide audience. Researchers, particularly in Natural Language Processing (NLP) and Information Retrieval (IR) domains, have seized this phenomenon, unprecedented by the number of users that these networks aggregate and the size of the data exchanged (texts, videos, audio...), opening up new research issues. Through these communication platforms, users can gather around a specific event (news [21], TV shows [25]...) which can even be recurrent (festivals [18], presidential elections [24]...).

The CLEF 2018 Mining opinion argumentation task aims to automatically identify messages of social web users positions about a cultural event expressed through the Twitter social network platform. The idea is to identify claims about a festival name, or topic, out of a massive collection of microblogs. The objective

^{*} This work was funded by the GaFes project supported by the French National Research Agency (ANR) under contract ANR-14-CE24-0022.

is to provide relevant information expressed in the form of a summary of argumentative tweets about a query (here a festival name or a topic) that should reflect a maximum of different points of view. This follows a previous task initiated in [7] about cultural microblog contextualization.

These last years, sentiment analysis and opinion mining [16] on social networks became an interesting field of study. Usually, many works proposed supervised approaches [12, 20] since annotated corpora are now available [15, 2]. Recent works have shown that convolutional neural networks (CNNs) are also well suited for sentence classification problems and can produce state-of-the-art results [23, 22, 19].

In this article, we propose an original four-steps approach to train a CNN model for extracting argumentative microblogs related to a specific query (or event) while no reference data is provided (and no data will be annotated).

The paper is organized as follows. Section 2 explains our proposed four-steps approach to identify a set of argumentative microblogs from a cultural event. Section 3 describes the experimental protocol, including a description of the task and the data used. Finally, Section 4 presents the results obtained in the CLEF 2018 Mining opinion argumentation task before concluding and exposing perspectives in Section 5.

2 Proposed Approach

In this section, we describe our proposed method to extract argumentative messages from a targeted query. Figure 1 summarizes our four-steps approach. The first step (Section 2.1) consists in preprocessing raw unlabeled messages to make them “cleaner”, *i.e.* make the data more easily interpretable and generalizable by an automatic process. The second step (Section 2.2) takes as input the cleaned data and proposes a method to extract two datasets (*Argumentative* and *Non Argumentative*) while no labeled data is provided. From these two datasets, a convolutional neural network (CNN) is trained in Step 3 to recognize argumentative and non argumentative messages (Section 2.3). Finally, the last step (Section 2.4) seeks to extract, from a set of messages related to a query (test set), the list of messages which contains the most argumentative elements while including a maximum of diversity in the opinions conveyed.

2.1 Preprocessing

In a general way, text messages need a preprocessing step to be then used as efficiently as possible in many NLP tasks. Usually, this process includes a global “cleaning” of the data. We first propose to tokenize words in order to better treat them individually. For example, the tweet “*It’s Friday, it’s Swansea Jazz Festival its cocktail night at Morgan’s.*” becomes “*It ’s Friday , it ’s Swansea Jazz Festival its cocktail night at Morgan ’s .*”.

Some specificities of tweet microblogs are also taken into account. Since URLs can be added in messages, we propose to make them unique by changing any

URL present in a tweet by $<URL>$. Nonetheless, as we think that hashtags ($\#example$) and references to other users ($@user$) are important information, we did not make any preprocess on it.

In many NLP applications [4, 9], word lemmatization seems to be a good way to improve performance. It regroups a family of words having different forms into a single form. For example, the words “learning” and “learned” will be grouped to “learn”, which should help by globally reducing the corpus vocabulary size. All datasets have been lemmatized with supervised part-of-speech taggers: LIA_TAGG³ and NLTK WordNet lemmatizers [3] for French and English messages respectively.

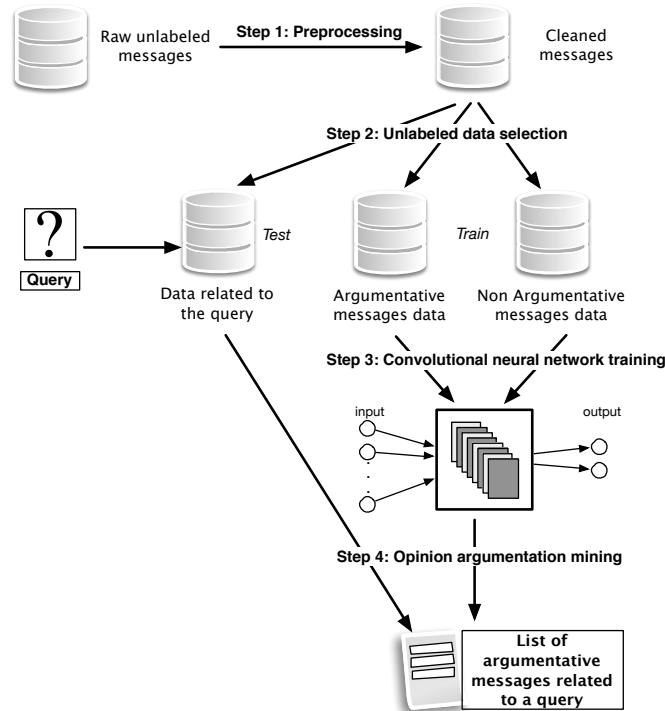


Fig. 1. Overview of the proposed approach for mining argumentative messages from raw unlabeled data regarding a targeted query.

³ <http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html>

2.2 Unlabeled data selection

While no reference data is available, we propose to “infer” this reference using a semi-supervised approach. For this unlabeled argumentative message data selection process, we firstly only keep messages tagged as the focused query language by the Twitter platform. Since this is an automatic process, errors in language identification may occur (and datasets may be different with another language identification tool).

We then get lists of opinion words that come from the French expanded emotion lexicon FEEL [1] (around 14k words) and an English opinion lexicon [10] (around 7k words).

Note that we do not define our training corpus only regarding these lists of opinion words: if so, it would simply amount to seeing the presence or absence of an opinion word to decide if a message is argumentative or not. We introduce new knowledge related to the corpus of microblogs studied: we make the hypothesis that a message can also be informative if it contains emoticons, particular punctuation signs such as ? or !, if the personal pronoun *Je* (in French) or *I* (in English) is employed, or if at least one hashtag is present. Possessive pronouns and personal pronouns are considered indicators of argumentative tweets for their expressive propriety. In particular first person and second person place the author in a communicational context expressive or conative.

In summary, we then have 5 features to decide if a message is informative (emotion words, emoticons, particular punctuation signs, personal pronoun, and hashtag). If a message contains at least 4 of these 5 features, or an emotion word plus 2 of the 4 other features, it is considered as argumentative. At the contrary, if a message does not have any of these characteristics, or only 1 feature (excluding an opinion word), it is considered as non argumentative.

This finally allows us to get two datasets for training: *Argumentative* and *Non argumentative*. Note that these train datasets have been extracted from all data excluding the data related to the targeted query, which constitutes here our database to search argumentative messages (*i.e.* test set). To constitute this test set, we consider a message related to a query if the words of the query are present in the message. For example, if the targeted query is “Avignon” for the French language, all messages containing the term Avignon and being tagged as French are in the test set, while all the remaining messages in the corpus (tagged as French) may be used to constitute the training data.

2.3 Convolutional neural network training

Convolutional neural networks (CNNs) represent one of the most used Deep Neural Network model in computer vision [13]. The difference between CNNs applied to computer vision and their equivalent in NLP lies in the input dimensionality and format. In computer vision, inputs are usually single-channel (eg. grayscale) or multi-channel (eg. RGB) 2D or 3D matrices, usually of constant dimension.

In sentence classification, each input consists of a sequence of words of variable length. Each word w is represented with a n-dimensional vector (word embedding) e_w of constant size. All the word representations are then concatenated in their respective order and padded with zero-vectors to a fixed length (maximum possible length of the sentence).

The parameters of our model were chosen so as to maximize performance on the development set (10% from the train data presented in Section 2.2): the width of the convolution filters is set to 5 and the number of convolutional feature maps is 200. We use ReLU activation functions and a simple max-pooling. One fully connected hidden-layers are of size 128. For each layer, a standard dropout of 0.4 (40% of the neurons are disabled in each iteration) is used. The back-propagation algorithm used for training is Adadelta.

2.4 Opinion argumentation mining

This last step allows us to constitute the list of argumentative message candidates. To do so, all the data test set (*i.e.* messages related to the query - see Section 2.2) is processed through the previously trained CNN. As a result, a score is assigned to each message that represents the probability of this message to be argumentative. A first ranked list can then be obtained with this classification process.

However, this first list does not respect the expected criterion of diversity of opinions: the list should reflect the maximum of argumentative points-of-view from a query (or event). In order to only keep enough different views, we compute a cosine similarity between a candidate message and the messages stored in this new list. Messages having a similarity higher than 0.5 are then excluded. For example, for the query *Rock festival* in English, if we get the following ordered list of candidate argumentative messages :

1. *common dave !! fuck the festival setting !! bless u with your awesome sitting acoustic rock ! ! ! ! #foofighters #pinkpop*
2. *managed to rock up in bordeaux on the weekend of both the gay pride festival and the main wine expo . #party*
3. *managed to rock up in bordeaux on the weekend of both the gay pride festival and the main wine expo*

The first message is automatically added to the final candidate list. Then, the cosine distance will be computed between the first and the second message: since they are different enough, the second message will also be added to the final list. For the third message, the cosine distance is computed with all the messages from the final list (messages 1 and 2): for the second message, the cosine distance is too close, the message 3 then does not finally appear in the final list of argumentative message candidates.

3 Experimental Protocol

The proposed approach has been assessed in the context of the CLEF 2018 Mining opinion argumentation task [8]⁴. A general description of this original task is proposed in Section 3.1, before describing the dataset in Section 3.2. Finally, Section 3.3 gives some details about the evaluation metric.

3.1 Task presentation

The general objective of the task is to find, for a specific topic or event, the most argumentative microblogs. These short messages come from a large collection of tweets about festivals in different languages. The idea is to get a list of ranked tweets, for each topic in a targeted language, according to their probability of being argumentative. Also, one key point lies in the opinion argumentation diversity provided in this list: a wide range of different points-of-view expressed in the tweets must be present (*i.e.* avoiding as much as possible identical argumentations).

This task may be of great interest to get a quick overview of opinions shared during an event from social networks since it is usually impossible to manually analyze all emitted messages. As a result, a set of 100 messages for each query (*i.e.* topic or event) must be given, each one being associated with a probability that the tweet is argumentative.

3.2 Data description

The CLEF 2018 Mining opinion argumentation task comes with a large collection of microblogs containing a stream of 70 million tweets in 134 different languages extracted from the Twitter platform. This dataset has been collected over a period of 18 months from May 2015 to November 2016 using a predefined set of keywords related to cultural festivals in the world [6]. Note that this 70 million tweet corpus includes the retweets⁵: if only the “original” posted messages are considered, the corpus is reduced to 33 million messages. In the proposed approach, the corpus considered is the one without retweets.

Regarding the targeted task, organizers propose to focus on two languages: French and English, from which 4 and 12 topics (*i.e.* queries) have been defined respectively. These queries have been chosen to match with festival names or topics. As explained by the organizers, these queries have enough related argumentative tweets to be evaluated. Table 1 lists these different topics or festival names for each considered language (French and English). For readability reasons, this list is presented in a descending order from the most popular topic (*i.e.* having the highest number of tweets) to the less popular one (*i.e.* having the smallest number of messages), each language being considered independently.

⁴ <https://mc2.talne.eu/>

⁵ A retweet is a forwarded message on Twitter. It is not an original post, but it is considered as a message.

Note that a message is linked to a festival name (or a topic) if it is present in the tweet content, no matter the language considered for now since there is no sure way (*i.e.* not automatic) to know the language of a tweet. The term *Festival* is excluded from this search since we assume that it is a Festival oriented corpus. An example of a tweet related to the *Cannes Festival*, where *Cannes* occurs:

At Cannes Film Festival, Dheepan Wins Palme dOr.

Table 1. List of queries (topic or festival name) for each considered language ordered by their number of messages (desc.) in the microblog corpus for the CLEF 2018 Mining opinion argumentation task.

Language	Query	# messages (all lang.)
French	Cannes Festival	1,470,882
	Rock Festival	1,232,529
	Jazz Festival	859,795
	Avignon Festival	55,109
English	Summer festival	1,715,017
	Cannes festival	1,470,882
	Rock festival	1,232,529
	Jazz festival	859,795
	Art festival	423,983
	Toronto festival	269,795
	Lantern festival	268,470
	Lollapalooza festival	133,111
	Texas festival	85,213
	Tomorrowland festival	66,176
	Bournemouth festival	21,057
	Hellfest festival	14,516

By analyzing Table 1 more precisely, we find that the festivals do not have the same activity as for the messages exchanged, with a huge difference between the most popular queries and the less popular ones. The *Cannes Festival*, which is the only festival name considered in both English and French languages, is the most represented in terms of posted messages. This is not surprising since it is a world famous festival. In the same way, the selected topics (*Rock*, *Jazz*, *Summer* and *Art*), chosen for being generic words, have a high level of activity, even if *Summer* appears well above others. Finally, the remaining festival names have the lowest number of tweets.

While these first observations may inform about the general corpus and this imbalanced queries data, Table 2 presents the dataset used for training our proposed system. Two subsets for training have been extracted for each query (*Argumentative* and *Non argumentative*). As expected, we find that many fewer tweets are annotated argumentative. A last subset, called *Test*, is composed of all the tweets containing the query. More information about this unlabeled argumentative tweet data selection process can be found in Section 2.2.

Table 2. Number of tweet messages in train and test datasets extracted from unlabeled data for each query. Queries (topics or festival name) are ordered by test dataset sizes (all query messages tagged as the targeted language).

Language	Query	Train		Test
		Argument.	Non argument.	
French	Cannes Festival	39,715	595,370	75,200
	Jazz Festival	42,063	628,109	19,098
	Avignon Festival	41,988	634,031	12,315
	Rock Festival	42,193	634,072	11,230
English	Art festival	300,752	7,781,925	329,172
	Summer festival	301,609	7,887,056	204,776
	Jazz festival	304,426	7,895,297	165,192
	Cannes festival	302,914	7,916,044	144,419
	Rock festival	304,275	7,958,201	92,818
	Toronto festival	305,541	7,981,868	52,625
	Lantern festival	306,172	8,000,335	27,368
	Texas festival	306,097	8,003,554	23,007
	Lollapalooza festival	306,555	8,013,735	7,800
	Bournemouth festival	306,607	8,015,118	5,111
	Tomorrowland festival	306,591	8,014,803	5,903
	Hellfest festival	306,650	8,018,223	1,166

Globally, we can firstly note that the imbalance in the data sizes (Table 2) is clearly reduced compared to Table 1. The *Cannes festival* remains the most commented festival name and the topics *Rock*, *Jazz*, *Summer* and *Art* datasets still have a high number of associated messages. Finally, for some festival names (especially for the English language), a very limited number of test data will be available, which make it difficult to get 100 argumentative microblogs.

3.3 Evaluation metric

The metric used to evaluate systems submitted to CLEF 2018 Mining opinion argumentation task is the Normalized Discounted Cumulative Gain (NDCG) [11]. It is a common ranking measure for IR tasks that gives a score for each retrieved argumentative tweet with a discount function over the rank. This measure takes into account the idea that the most interesting (*i.e.* argumentative) messages should appear first in the list while the non-relevant ones should not appear (or at the lower possible rank) [8]. Globally, the higher the measure is, the better the results are.

4 Results

Table 3 summarizes the results obtained by our system for the CLEF 2018 Mining opinion argumentation task in terms of NDCG score. For this task, two reference evaluation sets (*i.e.* sets of argumentative tweets) are considered: a manual

one, which corresponds to a fine manual annotation from the whole corpus, and a pooling one, which corresponds to a manual annotation from the tweets considered as argumentative by participants. For sake of comparison, three other systems are evaluated: the CLEF 2018 baseline [8], the LIA baseline (here, only spotting tweets considering opinion words) and the best system among all the CLEF 2018 participants.

Table 3. Performance, in terms of NDCG, of the proposed system (LIA_sub) on the manual reference. Two evaluation sets are considered: a manual one and a pooling one. Two baseline systems are also provided for comparison (LIA_base and CLEF_base) as well as the best performance (Best) from all the participants systems of the evaluation campaign.

Language	System	Manual ref.	Pooling ref.
French	LIA_base	2.886	0.150
	CLEF_base	2.285	0.049
	Best	2.894	2.057
	LIA_sub	2.894	0.067
English	LIA_base	0.061	0.047
	CLEF_base	0.007	0.173
	Best	0.061	0.601
	LIA_sub	0.061	0.063

By firstly focusing on the manual reference set, we can see that our proposed systems reach the best NDCG scores. Surprisingly, our baseline system (only opinion words) reaches similar results than our proposed system. This could be explained by the fact that opinion words may not be the only information to define what is an argumentative tweet. When focusing on the pooling reference, results are quite different: other participants systems reach much better performance. As a conclusion, we think that our system seems more robust regarding the whole corpus (best performance in the manual reference) by providing more diverse results than other participants (low performance in the pooling reference). All these observations are similar on French and English queries.

5 Conclusion and Perspectives

In this paper, the problem of retrieving argumentative microblogs from a large collection of messages was addressed. This work took place in the context of the CLEF 2018 Mining opinion argumentation task that aims to retrieve, for a specific event (festival name or topic), the most argumentative tweets from a Twitter festival-oriented corpus. To do so, we proposed an original CNN-based approach that takes into account the fact that no reference data is available (*i.e.* no tweets are annotated for training). As a result, a ranked list of the 100 most argumentative tweets, including an argumentative probability score for each message, has been provided for each query.

Results obtained on this evaluation campaign task appear encouraging, considering in particular the difficulty of the task. Indeed, our proposed approach reached best performance among all the participants on the manual reference. We also noted that this approach provides results very different from other participants, which has been observed on the pooling reference results. This could in particular open up perspectives of complementarity of the systems proposed for this evaluation campaign.

Many research perspectives can be gleaned from this preliminary work. Firstly, a robust language identification tool should be employed to select appropriate database. Another more interesting perspective would be to take account of the language level and particularity of tweet contents: indeed, microblogs exhibit particular linguistic characteristics (ungrammaticality, community-specific linguistic traits, misspelling...), not treated in this work. For example, a preprocessing method, such as [17], could be applied. These microblogs content particularities could also be treated with character-based approaches with adapted methods such as [5]. The use of the retweet information was also omitted in the proposed method. This information could be used in the selection process, for example by giving more importance to informative messages being very shared. Finally, it would be useful to explore methods in the field of automatic summarization that integrate the issue of content diversity, such as [14].

References

1. Abdaoui, A., Azé, J., Bringay, S., Poncelet, P.: Feel: a french expanded emotion lexicon. *Language Resources and Evaluation* **51**(3), 833–855 (2017)
2. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC. vol. 10, pp. 2200–2204 (2010)
3. Bird, S., Loper, E.: Nltk: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. p. 31. Association for Computational Linguistics (2004)
4. Carlberger, J., Dalianis, H., Duneld, M., Knutsson, O.: Improving precision in information retrieval for swedish using stemming. In: Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001) (2001)
5. Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., Cohen, W.: Tweet2vec: Character-based distributed representations for social media. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 269–274. Association for Computational Linguistics, Berlin, Germany (August 2016), <http://anthology.aclweb.org/P16-2044>
6. Ermakova, L., Goeuriot, L., Mothe, J., Mulhem, P., Nie, J.Y., SanJuan, E.: Clef 2017 microblog cultural contextualization lab overview. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 304–314. Springer (2017)
7. Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., SanJuan, E.: Overview of the clef 2016 cultural micro-blog contextualization workshop. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 371–378. Springer (2016)

8. Hajjem, M., Cossu, J.V., Latiri, C., SanJuan, E.: Clef 2018, international conference of the cross-language evaluation forum for european languages proceedings. In: LNCS volume, Springer, CLEF 2018, Avignon (France)
9. Halácsy, P., Trón, V.: Benefits of deep nlp-based lemmatization for information retrieval. In: CLEF (Working Notes) (2006)
10. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177. ACM (2004)
11. Järvelin, K., Kekäläinen, J.: Ir evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 41–48. ACM (2000)
12. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: The good the bad and the omg! Icwsm **11**(538-541), 164 (2011)
13. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361**(10) (1995)
14. Nema, P., Khapra, M., Laha, A., Ravindran, B.: Diversity driven attention model for query-based abstractive summarization. arXiv preprint arXiv:1704.08300 (2017)
15. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc. vol. 10 (2010)
16. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval **2**(1–2), 1–135 (2008)
17. Papegnies, E., Labatut, V., Dufour, R., Linarès, G.: Detection of abusive messages in an on-line community. In: 14ème Conférence en Recherche d'Information et Applications (CORIA). pp. 153–168 (2017)
18. Quillot, M., Ollivier, C., Dufour, R., Labatut, V.: Exploring temporal analysis of tweet content from cultural events. In: International Conference on Statistical Language and Speech Processing. pp. 82–93. Springer (2017)
19. Rouvier, M.: Lia at semeval-2017 task 4: An ensemble of neural networks for sentiment classification. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 760–765 (2017)
20. Rouvier, M., Favre, B.: Sensei-lif at semeval-2016 task 4: Polarity embedding fusion for robust sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 202–208 (2016)
21. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. pp. 851–860. ACM (2010)
22. Severyn, A., Moschitti, A.: Unitn: Training deep convolutional neural network for twitter sentiment classification. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado. pp. 464–469 (2015)
23. Tang, D., Wei, F., Qin, B., Liu, T., Zhou, M.: Coooolll: A deep learning system for twitter sentiment classification. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 208–212 (2014)
24. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In: Proceedings of the ACL 2012 System Demonstrations. pp. 115–120. Association for Computational Linguistics (2012)
25. Wohn, D.Y., Na, E.K.: Tweeting about tv: Sharing television viewing experiences via social media message streams. First Monday **16**(3) (2011)