Emerging Sentiment Language Model for Emotion Detection

Anastasia Giachanou[1], Francisco Rangel[2,3], Fabio Crestani[1], and Paolo Rosso[2]

[1]Faculty of Informatics, Università della Svizzera italiana (USI), Lugano, Switzerland
[2]PRHLT Research Center, Universitat Politècnica de València, Spain
[3]Autoritas Consulting, S.A., Spain
{anastasia.giachanou,fabio.crestani}@usi.ch
francisco.rangel@autoritas.es, prosso@dsic.upv.es

Abstract

English. In this paper we present an approach for joy, anger and neutral emotions detection based on an emerging sentiment language model. We propose an approach that can detect specific emotions from positive, neutral and negative sentiments and which favors the tweets that occur at recent sentiment spikes. Our results suggest that our approach can effectively detect joy, neutral and anger emotions and that it performs better compared to the baselines.

Italiano. *In questo articolo presentiamo un approccio per rilevare gioia, rabbia e emozione neutra basato su un modello di sentiment analysis emergente. Proponiamo un approccio in grado di rilevare emozioni specifiche da sentimenti positivi, neutri e negativi e che favorisca i tweets che si verificano nei picchi recenti di sentimento. I risultati suggeriscono che il nostro approccio può rilevare efficacemente le emozioni di gioia, rabbia e neutra che ottiene migliori risultati delle baseline.*

1 Introduction

Recent years have seen the emergence of social media that enable people to share their thoughts and opinions in an easy and fast way. Opinions posted on social media are very useful to understand what people think and how they feel about a specific entity (e.g. a product, a person, a company etc.). For example, companies can mine users' opinions on a product that has just been released to understand if the users are satisfied or not and act accordingly. Therefore, the automatic detection of emotions and sentiments from text has attracted a lot of research interest (Tang et al., 2015; Mohammad, 2015).

Although sentiment and emotion analysis share some similarities, they are two different problems (Munezero et al., 2014). *Sentiment analysis* focuses on understanding the sentiment polarity (positive, neutral, negative) of a text (Pang and Lee, 2008) whereas *emotion analysis* refers to its affectual attitude such as anger, joy, fear etc. (Mohammad, 2015). Most of the previous work have tried to predict sentiments from data annotated with sentiments and emotions from data annotated with emotions. However, preliminary experiments showed that some emotions are related to sentiments. Specifically, negative sentiment is related to anger, positive sentiment to joy and text with no emotion (neutral emotion) to text with no sentiment (neutral sentiment).

In addition, public sentiment towards a specific entity changes over time and in some cases sentiment spikes may occur. Sentiment spikes occur when a large amount of documents of a specific sentiment is posted (Giachanou et al., 2016). The documents that occur at sentiment spikes usually refer to a topic or event that attracted a lot of attention and therefore they can be very helpful for sentiment and emotion analysis. To this end, in this study we propose to incorporate information from the documents that have occurred at sentiment spikes to improve the performance of the emotion analysis task.

In this paper, we focus on Twitter and we propose the *emerging Sentiment Language Model (emerging-SLM)* approach which favors tweets that occur at recent sentiment spikes with the aim to predict joy, anger and neutral emotions from positive, negative and neutral sentiments respectively. We test our approach on a collection of tweets that spans over nine days and we show that the emerging-SLM performs better compared to both state-of-art Sentiment Language Model (SLM) and to a random Sentiment Language Model (random-SLM).

2 Related Work

Sentiment and emotion analysis have both attracted much research attention (Mohammad, 2015; Giachanou and Crestani, 2016). The main difference between the two problems is that sentiment refers to the polarity (e.g. positive, neutral, negative) whereas emotion refers to the affectual attitude that is anger, joy, fear etc. (Mohammad, 2015).

Sentiment analysis has attracted a tremendous research attention over the last years. The proposed approaches can be roughly classified as learning and lexicon based. The lexicon based approaches are typically unsupervised and use lists of words (e.g. *good*,*bad*) whose presence implies a specific sentiment polarity (Turney, 2002; Taboada et al., 2011). The learning based approaches rely on a number of features, usually extracted from text, to build a classifier which is then used to annotate unlabeled text as positive, negative or neutral (Pang et al., 2002). More recently, researchers have proposed deep learning approaches to learn sentiment specific word embeddings (Tang et al., 2014) or semantic representations of user and products (Tang et al., 2015) to address sentiment analysis. A thorough review on opinion retrieval and sentiment analysis can be found in Pang and Lee (2008) whereas Giachanou and Crestani (2016) focused on Twitter sentiment analysis.

With regards to emotion analysis, Mohammad (2012) considered hashtags that refer to an emotion (e.g., #anger, #surprise) to create a collection for emotion analysis and showed that these hashtag annotations matched with the annotations of trained judges. Roberts et al. (2012) extended the list of the six Ekman's basic emotions (joy, anger, fear, sadness, surprise, disgust) (Ekman, 1992) with an additional emotion (love) and created a series of binary SVM emotion classifiers. Also, other researchers have used sentiments or emotions to address other tasks such as irony detection (Farías et al., 2016) or author profiling (Rangel and Rosso, 2016).

In general, language models have been used for text classification problems (Bai et al., 2004). With regards to sentiment analysis, Liu et al. (2012) used manually annotated data to train a language model and then applied smoothing using noisy emoticon data. There are also few works that have considered sentiment dynamics. Bollen et al. (2011) used a psychometric instrument to extract and analyze different moods (tension, depression, anger, vigor, fatigue, confusion) detected in tweets and found that the mood level is correlated to cultural, political and other world global events while An et al. (2014) combined sentiment analysis, data mining and time series methods to track sentiment regarding climate change from Twitter feeds. However, our work is different because we use temporal information to favor documents that were posted recently and attracted a lot of attention with the aim to improve the performance of detecting specific emotions.

3 Methodology

Language Models (LMs) that are widely used in Information Retrieval (IR) and Natural Language Processing (NLP) fields assign probabilities to sequences of words (Ponte and Croft, 1998). The most typical scenario in IR consists in generating a Language Model (LM) for each document and then estimating the likelihood that the query was generated by each document. The documents then can be ranked based on the likelihoods. For a classification problem, we first aggregate all the documents of each specific class and then we estimate the likelihood that a new document is generated from each of the estimated language models. The new document can be annotated with the class for which it has the maximum likelihood.

More formally, let Θ^{+}, $\Theta^{\bullet}$, Θ^{-} be the LMs for the positive, neutral and negative classes respectively. Given a test tweet d we can detect its emotion class (joy, anger, neutral) c' as:

$$p(d|c') = \prod_{i=1}^{|d|} p(t_i|\Theta^c)$$

where $|d|$ is the number of words in tweet d and $p(t_i|\Theta^c)$ is a multinomial distribution estimated from the LM of class c (positive, negative, neutral).

To estimate the distributions we use the Maximum Likelihood Estimate (MLE) which computes the probabilities as follows:

$$p(t|\Theta^c) = \frac{n(t,c)}{\sum_{i=1}^{|V_c|} n(t_i,c)}$$

where $n(t,c)$ is the number of times that the term t appears in the collection of documents of class c and $|V_c|$ is the size of the vocabulary of class c.

The *emerging-SLM* combines two different LMs to estimate the probabilities of the terms. The first LM is based on all the tweets of the collection excluding those that occur at recent sentiment spikes whereas the second is based on tweets that occurred at those recent sentiment spikes. Formally, the distributions of the terms using the *emerging-SLM* are estimated as:

$$p(t|\Theta^c) = \lambda * p_{global}(t|\Theta^c) + (1-\lambda) * p_{burst}(t|\Theta^c)$$

where Θ^c is the LM for the class c, $p_{burst}(t|\Theta^c)$ is the probability of that term t appear in the recent sentiment spikes of the class c, $p_{global}(t|\Theta^c)$ is the probability of that term t appear in the class c and λ is the parameter that determines the importance of each LM for the final estimation. Here we should note that $p_{global}(t|\Theta^c)$ is calculated after we excluded the tweets that occurred at sentiment spikes.

One common issue with the LMs is that they assign zero probabilities to terms that do not appear in the training data. To overcome this problem, we apply Jelinek-Mercer smoothing that assigns nonzero probabilities to unseen terms (Zhai and Lafferty, 2004). Jelinek-Mercer smoothing refers to a linear interpolation of the MLE and the collection language model $p(t|\Theta_c)$ and can be defined as:

$$p(t|\Theta) = \mu * p(t|\Theta) + (1 - \mu) * p(t|\Theta_c)$$

where the collection language model is estimated using the maximum likelihood estimate of the whole collection.

To detect the sentiment spikes, we measure the evolution of each sentiment as $r_{t,s} = N_{t,s}/N_t$ where $N_{t,s}$ is the number of documents that express the sentiment s posted at time t and N_t is the total number of documents posted at time t. Figure 1 shows an example of negative spikes that occurred while tracking the sentiment towards *Michelle Obama*.

4 Experimental Setup

In this section we describe the experimental details of our study that include the description of the dataset, the baselines we used and the experimental settings.

4.1 Dataset

Our collection contains 25,588 tweets about *Michelle Obama* and spans from June 25, 2015 to July 2, 2015. To annotate the collection, we used the Crowdflower platform[1]. Tweets were annotated with regards to sentiment and emotion. For sentiments, annotators could choose among {*positive, no sentiment, negative*} whereas for emotions they could choose among {*anger, fear, sadness, disgust, surprise, happiness, no emotion*}. Each tweet was annotated by three different workers.

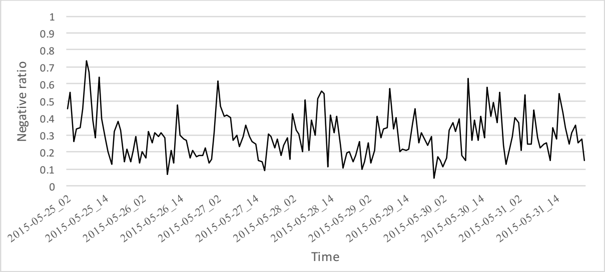

Figure 1: Negative spikes that occurred while tracking the sentiment towards *Michelle Obama*.

To optimize the annotation process and obtain more labels we applied a type of distant supervision, which is a popular technique for obtaining more labels for the data (Go et al., 2009). In our study we used the similarity between the tweets because a large amount of tweets are posted again (retweets). Therefore, first we ranked the tweets by how may times they were retweeted and then we collected annotations for the most popular ones. Next, we disseminated the labels to the rest of the tweets using a similarity threshold set to 0.8. We used cosine similarity to measure the similarity between two tweets.

For all the results reported to this study, we used 429 tweets as a test set which were posted on July 2, 2015. We kept the test and training data always separated.

4.2 Baselines

We used two different baselines to compare the performance of our approach. The first baseline (SLM) is based on sentiment language models and was built from all the data without favoring tweets that occurred at spikes (i.e. $\lambda = 1.0$). The second baseline is the random-SLM approach. In this case, instead of using tweets from recent sentiment spikes, we randomly chose tweets from the whole collection. To build the random LM we select as many tweets as those used to build the LM of sen-

[1] `https://www.crowdflower.com/`

timent spikes. To evaluate the statistical significance of differences we used the McNemar test.

4.3 Experimental Settings

For pre-processing, we removed URLs, mentions, punctuation and the entity-related terms *Michelle* and *Obama*. For the experiments we used only unigrams. To overcome the problem of assigning zero probabilities to unseen terms we used Jelinek-Mercer smoothing with $\mu = 0.1$.

To model the evolution of sentiment and detect any sentiment spikes we split the data hourly. In addition, we defined temporal bins with the size of 8 hours. For the *emerging-SLM* we detected all the sentiment spikes that occurred in the last two days. To detect the spikes, we used the peakutils[2] package setting the threshold to 0.8.

Finally, to tune the λ parameter, we used cross-validation on a rolling basis. Following this approach, we used data published on the first temporal bin as training and data of the second temporal bin as test. Next, data from the first and second temporal bins were used for training and data from the third temporal bin as test and so on. In other words, when we set the number of bins to 9, it means that we were using 3 days as training data (i.e. 8 hours * 9 bins = 72 hours) and one bin as test data. The test bin was always the adjacent temporal bin. The first setup included 3 days, since we wanted to have enough data to build the SLMs. After this process we estimated the best λ parameters using the average performance.

5 Results

Figure 2 shows the performance with regards to the F1-measure for the task of emotion detection using the emerging-SLM on the training data for the different parameters of λ. We show the results for six different temporal bins for reasons of clarity. From this figure, we observe that there is a performance improvement as the λ parameter increases.

Table 1 shows the performance with regards to F1-measure for the task of emotion detection using the emerging-SLM, the SLM and the random-SLM approaches. From the results we observe that the emerging-SLM performs better compared to SLM and random-SLM for all the temporal bins. Also, most of the differences are significant. These results are very important because they show that favoring tweets that have occurred at recent sentiment spikes is very useful. Also, the improvement over the random-SLM validates further this assumption. The results are also shown on Figure 3 for an easier comparison.

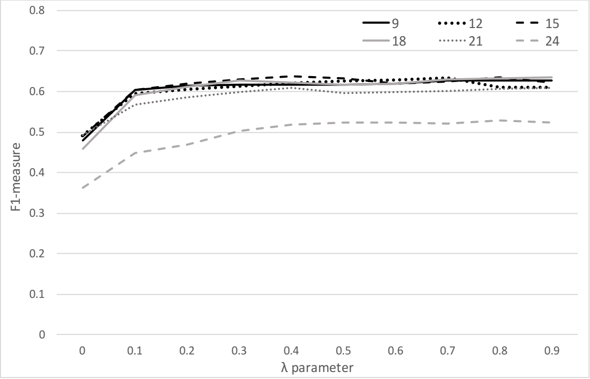

Figure 2: Performance of the *emerging-SLM* on the training data with regards to F1-measure for different λ parameters for six different bins.

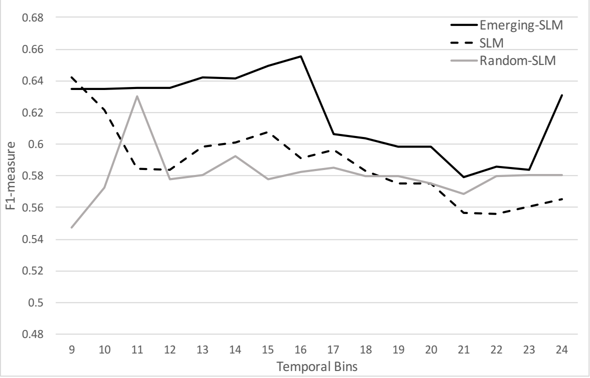

Figure 3: Performance measure with regards to F1-measure using different temporal bins.

6 Conclusions and Future Work

In this paper, we proposed the emerging-SLM approach to detect joy, neutral and anger emotions from positive, neutral and negative sentiments respectively. Emerging-SLM favors tweets that occur at recent sentiment spikes. The results showed that our approach performs better compared to both SLM and random-SLM and can be effectively applied to detect specific emotions.

In future we plan to explore if there is any effect of the temporal bins size on the emotion detection performance and if sentiment language models can be used to detect also other emotions such as fear and surprise.

[2] https://bitbucket.org/lucashnegri/peakutils

Table 1: Performance results over the test data using different size for the temporal bins of the emerging-SLM, SLM and random-SLM approaches. A star (∗) means there is a statistically significant difference between the emerging-SLM and SLM ($p<0.05$). A † indicates a significance difference between the emerging-SLM and random-SLM ($p<0.05$).

Bins	emerging-SLM	SLM	random-SLM
9	0.6352†	0.6419	0.5473
10	0.6352∗†	0.6213	0.5725
11	0.6358∗†	0.5842	0.6303
12	0.6358∗†	0.5841	0.5781
13	0.6419∗†	0.5980	0.5803
14	0.6415∗†	0.6012	0.5922
15	0.6493∗†	0.6078	0.5780
16	0.6551∗†	0.5914	0.5821
17	0.6064∗†	0.5961	0.5853
18	0.6034∗†	0.5828	0.5798
19	0.5987∗†	0.5751	0.5797
20	0.5987∗†	0.5751	0.5750
21	0.5792∗†	0.5564	0.5683
22	0.5855∗†	0.5557	0.5796
23	0.5837∗†	0.5604	0.5804
24	0.6311∗†	0.5651	0.5805

Acknowledgement

The work of the first author was partially funded by the Swiss National Science Foundation (SNSF) under the project OpiTrack.
The work of the last author was in the framework of the Spanish MINECO research project SomEMBED (TIN2015-71147-C2-1-P).

References

Xiaoran An, R. Auroop Ganguly, Yi Fang, B. Steven Scyphers, M. Ann Hunter, and G. Jennifer Dy. 2014. Tracking climate change opinions from twitter data. In *Proceedings of the Workshop on Data Science for Social Good held in conjunction with KDD 2014*.

Jing Bai, Jian-Yun Nie, and François Paradis. 2004. Using language models for text classification. In *Proceedings of the Asia Information Retrieval Symposium, Poster Session*, AIRS '04.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 450–453.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3/4):169–200.

Delia Irazú Hernańdez Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology*, 16(3):19:1–19:24.

Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys*, 49(2):28:1–28:41.

Anastasia Giachanou, Ida Mele, and Fabio Crestani. 2016. Explaining sentiment spikes in twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2263–2268.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Standford.

Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. 2012. Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 1678–1684.

Saif M. Mohammad. 2012. #Emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation*, SemEval '12, pages 246–255.

Saif M. Mohammad. 2015. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion measurement*, pages 201–238.

Myriam D. Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, EMNLP '02, pages 79–86.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281.

Francisco Rangel and Paolo Rosso. 2016. On the impact of emotions on author profiling. *Information Processing & Management*, 52(1):73 – 92.

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on twitter. In *Proceedings of the 8th International Language Resources and Evaluation Conference*, LREC '12, pages 3806–3813.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 1555–1565.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL '15, pages 1014–1023.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.