Distant-supervised Language Model for Detecting Emotional Upsurge on Twitter

Yoshinari Fujinuma[†,‡*] **Hikaru Yokono**[‡] **Pascual Martínez-Gómez**[§,‡] **Akiko Aizawa**[†,‡]
[†]University of Tokyo [‡]National Institute of Informatics [§]Ochanomizu University
fujinumay@gmail.com {yokono, pascual, aizawa}@nii.ac.jp

Abstract

Event-specific twitter streams often reveal sudden spikes triggered by users' upsurge of emotions to crucial moments in the real world. Although upsurge of emotion is usually identified by a sudden rise in the number of tweets, the detection for diverse event streams is not a trivial task. In this paper, we propose a new method to extract spiking tweets with upsurge of emotions based on characteristic expressions used in tweets. The core part of our method is to use a distant-supervised language model (Spike LM) built from tweets in spikes to capture such expressions. We investigate the performance of detecting emotional spiking tweets using language models including Spike LM. Our experimental results show that the natural language expressions used in emotional upsurge fit specifically well to Spike LM.

1 Introduction

Twitter is one of the most popular micro-blogging platforms in recent days. There are over 500 million tweets posted per day[1] including real-world events described on Twitter which range from short and daily life events (e.g. falling to the ground) to long and widely-broadcasted events (e.g. a match in World Cup). Such tweets are good sources to detect users' reactions toward real-world events.

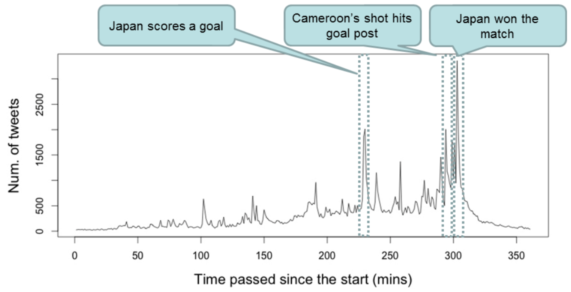


Figure 1: The number of tweets per minute (TPM) during Japan vs. Cameroon for hashtags related to World Cup 2010.

People behave unusually when they encounter exciting moments in an event, for example, yell out or dance with each other after their favorite soccer team scores a goal. On Twitter, this action is often reflected by a large number of posts within a short time period. When Japan scored a goal against Cameroon in World Cup 2010, there were a maximum of $2,940$ tweets per second (TPS), which marked the record TPS for goals at that time.[2] It is significantly larger than the average of 750 TPS.[2] In this paper, we call such bursty traffic as "numerical spikes". Figure 1 shows the number of tweets per minute during the match of Cameroon vs. Japan, and Table 1 shows the examples of tweets sampled from both numerical spikes and other parts.

Detecting emotional upsurge is important for both extracting emerging important real-world events and important moments of them. We call an upsurge that are caused by Twitter users' emotional spike as

[*] This work was done while the first author was at the University of Tokyo and National Institute of Informatics. The first author is currently at Amazon Japan K.K.

[1] https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

[2] https://blog.twitter.com/2010/big-goals-big-game-big-records

29th Pacific Asia Conference on Language, Information and Computation pages 488 - 495
Shanghai, China, October 30 - November 1, 2015

	Moment	Example of a Tweet	English Translation
Emotional Upsurge	Japan won the match	やったああああああああああああああああああああ #jpn #worldcup #2010wc	Huraaaaaaaaaaaay
	Japan scored a goal	ゴーーーーーーーーーーーーーーーーーーーール !!!!!!!! #2010wc	Gooooooooooooal!!!!!!!!
Non-emotional Upsurge	50 mins after the match	興奮してると見せかけて感動しすぎてずっと泣いてました。いや興奮はしてるけど信じてたから割と冷静でいられる #2010wc	I look excited but actually I have been crying from being moved. Well, I have been excited but I believed that Japan will win so I am quite calm.

Table 1: Example of tweets from spikes and non-spikes.

"emotional upsurge". Emotional upsurge do overlap with numerical spikes, but it does include moments that are not numerical spikes. For example, Lanagan and Smeaton (2011) reported that emotional upsurge overlaps with numerical spikes and those are useful for tagging key moments in sports matches. However, detecting numerical spikes on Twitter becomes difficult when a target event is not pre-defined or rarely tweeted by Twitter users because the number event-related tweets per unit time is not directly computable. In such cases, detecting upsurge of emotions becomes crucial.

One characteristics of tweets is that expressions used in tweets entail many linguistic phenomena. For example, Brody and Diakopoulos (2011) analyzed occurrences of character repetitions in words from a sentiment dictionary. In this paper, we assume that such variations of language expressions are caused by real-world events. Table 1 shows that a character repetition ('Goooal', 'Huraaay') occurs in tweets during emotional upsurge rather than their canonical form ('Goal', 'Hurray'). In contrast, a character repetition does not frequently occur in tweets during non-emotional upsurge. However, to our knowledge, there has not been an attempt to capture emotional upsurge using the linguistic characteristics of tweets.

In this paper, we specifically investigate a method to detect emotional upsurge in real-world events using characteristic expressions in a Japanese tweet. Our contribution is that a *spiking tweet language model*, which we constructed automatically from existing tweet dataset, captures characteristic expressions well and it is an effective approach for detecting emotional upsurge.

2 Related Work

Our idea is related to many previous works on Twitter including the investigation toward non-standard languages used on Twitter, and various applications tackled using language models.

The nature of using non-standard languages including word lengthening in tweets largely differ from other corpus (Eisenstein, 2013). As further mentioned by Eisenstein (2013), these languages are affected by many factors including the 140 characters length limit of tweets, social factors (e.g. age (Rosenthal and McKeown, 2011)), location (Wing and Baldridge, 2011), input devices (Gouws et al., 2011) of an author of a tweet. Word lengthening is known to be useful for sentiment analysis (Brody and Diakopoulos, 2011). One way to model these expressions is to use language models and many studies successfully captured various characteristics of tweets using language model.

There are lots of applications for language models built from tweets or web texts. According to Liu et al. (2012), distant-supervised language models are useful for sentiment analysis of tweets. Neubig and Duh (2013) showed that for 26 languages used on Twitter, entropy of content in a retweet, the Twitter version of e-mail forward, is significantly higher than non-retweeted tweets. Danescu-Niculescu-Mizil et al. (2013) reported that users' career in an online community correlates with the cross entropy between each user's posts and the language used in the whole community. Lin et al. (2011) used multiple language models built from each hashtag to track broad topics. These researches show that language model is a powerful method to use on various applications.

To our knowledge, there is no prior research focused on languages used in emotional spiking tweets. Many tasks on Twitter including burst detection (Kleinberg, 2003; Diao et al., 2012), first story detection (Petrović et al., 2010), and topic tracking (Lau et al., 2012) failed to effectively incorporate the textual characteristics of tweets and regard it is out of their scope. Being able to characterize tweets from emotional upsurge would open a window to the identification of real-world events that emotionally influence Twitter users.

3 Language Model-based Detection of Emotional Upsurge

3.1 Outline of the Proposed Method

The motivation of using characteristic expressions used in tweets to detect emotional upsurge is there are various ways to express users' feelings. In the past investigations (for example (Schröder, 2001)), the emotion of a human speaker reflected by the tone or the pitch of speaker's voice. On Twitter, Brody and Diakopoulos (2011) reported that word lengthening in written words is used to express the difference in such user's voice, which is affected from user's sentiments. As shown earlier in Table 1, we assume that the language used in tweets can express difference in pitch or tone of voice as a written text in tweets. Therefore, we aim to capture such difference in voice-reflected tweets using language models.

In our approach, we further apply a distant supervision framework where the perplexity is calculated using a language model obtained from tweets in numerical spikes with some heuristic filtering strategy. If the perplexity of target tweets is small, we could then assert that they are likely to have come from the emotional tweet model.

3.2 Building Language Models

Since we can obtain a large number of tweets, we build tweet language models such that the language model is not biased by a particular topic. Given a tweet t with l characters, let t_i be a character in a tweet. The probability of t in an n-gram language model is calculated by the following formula:

$$P(t) = \prod_{i=1}^{l} P(t_i | t_{i-1}, ..., t_{i-n+1}). \qquad (1)$$

We use SRILM (Stolcke, 2002) with Katz back-off smoothing (Katz, 1987) to build language models.

We build a character n-gram language model following Neubig and Duh (2013). To build a word n-gram language model, word segmentation is necessary to build a word n-gram language model since Japanese is an unsegmented language. However, various studies reported that tokenization in unsegmented languages on Twitter is not reliable enough due to the spelling variations and unknown words (Wang and Kan, 2013; Kaji and Kitsuregawa, 2014). We set the value of n for a character n-gram language model to 7. This is because when we consider n-grams with $n > 5$, the number of n-grams decreases which shows that the language model suffers from the sparsity problem. However, as reported by Brody and Diakopoulos (2011), word lengthening (e.g. coool) is a common phenomenon on Twitter. To accurately capture those phenomenon, we tried to use as long n-gram as possible and make it to 7-gram.

3.3 Perplexity

To quantify the difference between tweets during emotional upsurge and non-emotional upsurge, we used perplexity, a measurement of information-theoretic distance between a language model and a document. In this method, it is used as the similarity between a language model and a set of tweets. Perplexity PP of a tweet set T which consists of N number of 7-grams T_i is defined as the following:

$$PP(T) = \left(\frac{1}{\prod_{t \in T} P(t)} \right)^{\frac{1}{N}}. \qquad (2)$$

Hashtag	Details	Date and Time	Num. of Tweets	Num. of EUT
#aibou	Name of a TV drama	2012-03-21T10:31 - 14:06	20,681	42
#hanshin	Name of a baseball team	2012-04-20T08:39 - 12:54	6,176	44
#ACV	Name of an online game	2012-02-13T12:08 - 15:41	1,562	9
#agqr	Name of a radio show	2012-02-15T11:39 - 14:08	13,434	86
#figureskate	Figure skating	2012-04-20T10:00 - 12:20	1,410	37
#momoclo	Name of a music artist	2012-02-11T15:36 - 17:21	1,823	63

Table 2: Statistics of six hashtags, its respective target intervals and with the number of manually annotated emotionally upsurging timestamps (EUT). All time are UTC +0.

We follow Danescu-Niculescu-Mizil et al. (2013) and use perplexity, which is equivalent to the cross-entropy of two empirical distributions, as similarity measure between a set of tweets and a language model. A set of tweets having low perplexity means tweets are close to the language model. Moreover, we assume that the minimum duration of an event is a minute. Therefore, we aggregate stream of tweets from the same minute into one set of tweets and compute the perplexity of it.

4 Tweet Dataset Construction

The dataset we use in this paper is extracted from Japanese tweets gathered by Gnip[3] during 5 consecutive periods such as 1) 2012-02-09 to 2012-02-17, 2) 2012-03-21 to 2012-03-22, 3) 2012-04-20 to 2012-04-21, 4) 2012-05-18 to 2012-05-19 and 5) 2012-05-25 to 2012-05-26. The dataset has a total of $413,008,939$ tweets with $527,661$ unique hashtags.

We construct four separate sub-datasets each of which is used for different purpose: evaluating the performance of detecting upsurge of emotion, building a language model. Since our research focuses on users' reactions to events, we filtered out tweets from bots[4] (Twitter accounts that automatically produce tweets according to a program) and tweets that include the characters 'RT', which indicates a retweet. For constructing the language models and the evaluation data, we regarded the Twitter specific elements such as hashtags, users (e.g. @Obama), hyperlinks as one character.

Dataset Used for Evaluating Language Models

To build a golden dataset of emotionally upsurging timestamps, we first extract hashtags which consist of both emotional upsurge and non-emotional upsurge from various genres. First, we selected six hashtags and we set a target interval for each hashtag as consecutive periods with the number of tweets ≥ 10 together with 20 minutes before and after the periods.

Next, we randomly sample 90 minutes from the target interval of each hashtag and aggregated tweets from the same minute as one tweet set. An annotator looks at all of the tweets from the same minute and annotate whether each timestamp is an emotional upsurge or not. As a result, 42 timestamps in #aibou, 44 timestamps in #hanshin, 9 timestamps in #ACV, 37 timestamps in #figureskate, 86 timestamps in #agqr and 63 timestamps in #momoclo are annotated as emotionally upsurging timestamps. Table 2 shows the details of the six hashtags and the target respective interval we used.

Dataset used for Building Spike LM

In order to construct a spiking tweet language model (Spike LM), we gather $1,197,935$ tweets

[3] http://gnip.com/

[4] Bots typically tweet from particular Twitter clients; thus, by looking at sampled data, we chose to use tweets from the top 43 Twitter clients in terms of frequency. These are not bots and covered over 90% of the tweets that we sampled for 3 days.

from all hashtags which exceed 50 TPM. We filter out hashtags including the word “follow” or “Follow” due to the large number of Twitter-specific hashtags. Moreover, we also exclude the six hashtags described in Table 2.

Dataset used for Building Supervised LM

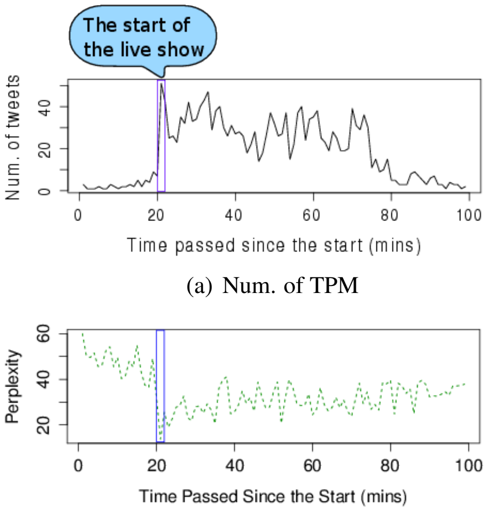
The limitation of the Spike LM is that we cannot avoid including tweets not from emotional upsurge. We build a fully supervised spike language model (Supervised LM) to observe whether clean but much low number of tweets will perform better than the language model built from less cleaned but more number of tweets. In order to filter out hashtags that include non-emotional numerical spikes, we used manually annotated emotionally upsurging timestamps (EUT) shown in Table 2, and constructed Supervised LM excluding the hashtags used for testing.

Dataset used to Evaluate Detecting Numerical Spikes using Spike LM

Since numerical spikes has some overlaps with emotional upsurge, we analyze if Spike LM can also detect numerical spikes. To analyze the detection of numerical spikes using Spike LM, we construct a tweet set containing 300 tweets from each hashtag in Table 2. The tweet set consist of one tweet set of 150 tweets sampled from numerical spikes and another tweet set of 150 tweets sampled from non-numerical spikes.[5] We compute the perplexity of a tweet set rather than to individual tweets to get a reliable perplexity.

5 Evaluation of Language Models

We evaluate how well does Spike LM detect numerical spikes and then compare the performance of detecting emotional upsurge against Supervised LM and Kleinberg’s algorithm.



(a) Num. of TPM

(b) Spike LM

Figure 2: Number of tweets in #momoclo hashtag shown together with the perplexity between the three language models. The blue box represents an example of a spike.

5.1 Evaluation of Detecting Numerical Spikes using Spike LM

We first evaluate the effectiveness of capturing numerical spikes on Spike LM. Figure 2 shows the perplexity of spiking timestamps are actually low compared to other timestamps. As a quantitative comparison, we sampled 300 tweets from each hashtag and calculate the perplexity of both numerical spiking and non-numerical spiking tweet sets using Spike LM as mentioned earlier in Section 4. Table 3 shows that for all the six hashtags, there is a significant difference between the perplexity of numerical spiking and non-numerical spiking tweets according to the Wilcoxon signed-rank test ($p < 0.02$). Therefore, Spike LM is useful for detecting tweets from numerical spikes.

Next, we evaluate the performance of detecting emotional upsurge from tweet sets aggregated by its timestamps using language models. We use the manually annotated ground truth emotional upsurge for evaluation.

5.2 Evaluation of Detecting Emotional Upsurge

To evaluate which language model best detect emotional upsurge, we derive precision, recall and F1-score for each language model by incrementing the perplexity decision threshold one by one. Specifi-

[5] To avoid test sets being biased from one incident, we constructed the test sets as the following steps: 1) Split the target interval into 3 sub-intervals. 2) For each sub-interval, gather the tweets from the most tweeted minutes until the total of number of tweets reaches 50. 3) If the number of tweets during the minutes exceeds 50, randomly sample 50.

Hashtag	PP(S)	PP(NS)	PP(NS)-PP(S)
#aibou	22.027	27.735	5.708
#hanshin	30.705	63.416	32.711
#ACV	43.116	52.647	9.531
#agqr	9.938	23.134	13.196
#figureskate	27.505	39.176	11.671
#momoclo	23.261	39.283	16.022

Table 3: Perplexity of sampled set of tweets constructed from numerical spikes (S) and non-numerical spikes (NS) computed using Spike LM.

Hashtag	Spike LM	Sup LM	Kleinberg
#aibou	.656	.655	**.667**
#hanshin	**.707**	.642	.508
#ACV	**.571**	.500	.400
#agqr	**1.00**	**1.00**	.491
#figureskate	**.643**	.595	.500
#momoclo	.527	**.817**	.615

Table 4: The best F1-score of detecting annotated emotional upsurge for Spike LM, Supervised LM (Sup LM) and Kleinberg's algorithm.

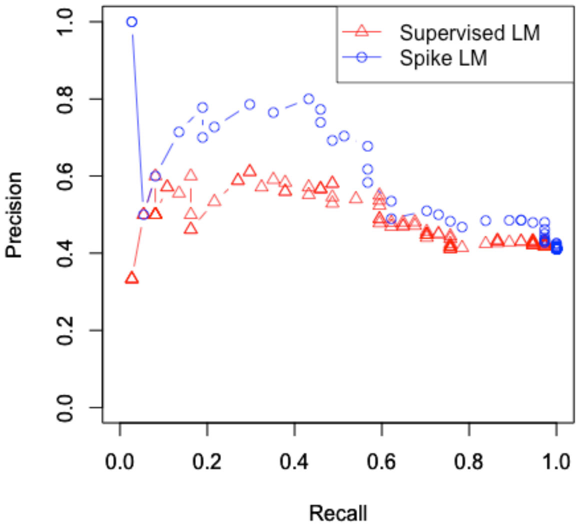


Figure 3: Precision-Recall curve for #figureskate data.

cally, if the perplexity of a set of tweets from one timestamp is lower than a decision threshold, we predict as a timestamp of emotional upsurge and vice versa. We eventually use the best F1-score among the various decision thresholds to evaluate which language model best fits to modeling emotional upsurge.

Baseline System

We employ Kleinberg's burst detection algorithm (Kleinberg, 2003) as a baseline method. This method assumes that all numerical spikes or bursts are emotional upsurge and all non-numerical spikes or non-bursts are not emotional upsurge. Kleinberg's burst detection algorithm modeled a burst of a stream of documents as a two-state finite state automata $B_{s,\gamma}$ with the scaling parameter s and the transition cost parameter γ. The states are assumed to be in either the burst state or the non-burst state. We further choose the optimal state sequence that requires minimum cost among all possible state sequences. As a result, we detect which timestamps are in a burst states and which timestamps are not. The two parameters of the algorithm is set according to the result from the preliminary experiment to detect numerical spikes based on mean and standard deviation with sufficiently high F1-scores. Specifically, the parameters γ and s are set to $\gamma = 1$ and $s = 2$.

Evaluation Result

Table 4 shows the result of detecting emotional upsurge for the two language models and Kleinberg's algorithm. Figure 3 shows the precision-recall curve for the two language models we built. According to this figure, Spike LM performs well among the majority of the test hashtags when compared to Supervised LM. Furthermore, the figure shows that the precision of Spike LM does not drop when we increase the decision threshold.

We observe that if a language model contains hashtags with similar emotional upsurge to the test hashtags, the performance of detecting emotional upsurge tend to get better. This is obvious for Supervised LM performing well on #momoclo hashtag because when testing on this hashtag, Supervised LM is built from the rest of five target hashtags including #agar and the suffix of the emotionally upsurging tweets from #momoclo are similar to that of #agqr. Specifically, those tweets include lots of "w"s which is an Internet slang meaning “lol (laugh out loud)” in English. Note that the effect of #agqr is magnified on Supervised LM since most of the annotated timestamps are annotated as emotional upsurge in #agqr.[6] This also explains why Spike LM performs better than Supervised LM on five hashtags because Spike LM is more likely to include emotional upsurge from hashtags similar to the six hashtags.

One of the challenges is to detect emotional upsurge with relatively low number of tweets because of the existence of noisy tweets. Example of noisy tweets are the tweets from Twitter accounts that only post about news. Table 5 shows an emotional spike including such noisy tweets, which scored 42.095 as the perplexity. This spike includes 7 tweets from #figureskate hashtag. This is relatively low since the number of tweets per timestamp in sampled #figureskate tweets range from 2 to 50. Among the 7 tweets, 2 tweets are from a Twitter account which only tweets about news, which does not reflect emotional upsurge of a Twitter user. Spike LM is robust to such noisy 2 tweets from the account which only tweets about news when computing the perplexity of that timestamp. However, Supervised LM is largely affected by such noisy tweets because Supervised LM is built from less noisy tweets compared to Spike LM and it end up with high perplexity. Spike LM detects such emotional upsurge which can be used to extract emotional upsurge from various domains on Twitter.

6 Discussion

We further investigate the impact of the tweet set size on the reliability of the perplexity estimation using language models. Perplexity is known to be affected by the amount of text used for the calculation (Brown et al., 1992). We analyzed the impact using the most tweeted minute in the hashtag #aibou. Figure 4 shows the transition of perplexity according to the number tweets used to calculate the perplexity in the hashtag #aibou. As a result, after the number of tweets from the same minute exceeds 11, the difference between the minimum and the maximum perplexity became less than 3. This result shows that the perplexity does not largely rely on the number of tweets from the same timestamp and implies that Spike LM can be used to detect emotional upsurge with low number of tweets.

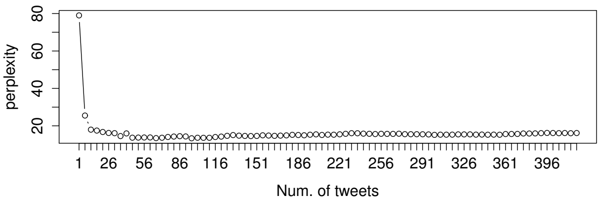


Figure 4: Perplexity computed with various number of tweets from the most tweeted minute in #aibou.

7 Conclusion

In this paper, we showed that sequences of tweet characters in emotional spiking tweets are more similar to that of tweets modeled by Spike LM. By calculating the perplexity between Spike LM and sampled tweets from numerical spikes and non-numerical spikes among multiple hashtags, tweets from numerical spikes had lower perplexity than tweets from non-numerical spikes. Furthermore, Spike LM scored the highest F1-scores for detecting emotional upsurge in over half of the hashtags we examined. In conclusion, our method detects tweets that include Twitter users’ upsurge of emotions, without largely depending on the number of tweets per minute by seeking for tweets modeled by Spike LM.

As a future task, we plan to investigate three further points: 1) Applying our method to other events tweeted on Twitter, 2) classification of emotional upsurge and non-emotional upsurge on the tweet level since we only investigated on a tweet set level, and 3) Test it on languages other than Japanese. Further studies are necessary to capture emotional spiking tweets on Twitter.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 22240007. The authors would like to thank the anonymous reviewers for their constructive comments.

[6] Therefore, both language models score 1.0 on #agqr.

Tweets	English Translation
ジュベ様ジュベ様!! うわああああん大好きいい！衣装しっかりいいい #figureskate	Joubert! Joubert! Ahhhhhhh, I love him! His outfitsssss
ジュベール様のマトリックス (Happy Emoticon) #figureskate	Joubert's Matrix (Happy Emoticon)
男子シングルの結果です。→ URL #figureskate 高橋大輔選手が圧巻の演技を見せ、世界王者・チャンを破り1位に。日本チームとしてもトップのまま、最終日を迎えます。 #figureskate	This is the result of the mens' figure skating competition. → URL Daisuke Takashi showed an amazing performance and he became the number one after beating the world champion Chan. He is also the number one as part of the Japan team and reach the final day.

Table 5: An example of detected emotional upsurge with low number of tweets per minute.

References

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *EMNLP*, pages 562–570.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *WWW*, pages 307–318.

Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *ACL*, pages 536–544.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *NAACL-HLT*, pages 359–369.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proc. of the Workshop on Language in Social Media (LSM 2011)*, pages 20–29.

Nobuhiro Kaji and Masaru Kitsuregawa. 2014. Accurate word segmentation and pos tagging for japanese microblogs: Corpus annotation and joint modeling with lexical normalization. In *EMNLP*, pages 99–109.

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 400–401.

Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

James Lanagan and Alan F Smeaton. 2011. Using twitter to detect and tag important events in live sports. *Artificial Intelligence*, pages 542–545.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line trend analysis with topic models: #twitter trends detection topic model online. In *COLING*, pages 1519–1534.

Jimmy Lin, Rion Snow, and William Morgan. 2011. Smoothing techniques for adaptive online language models: Topic tracking in tweet streams. In *KDD*, pages 422–429. ACM.

Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. 2012. Emoticon smoothed language models for twitter sentiment analysis. In *AAAI*, pages 1678–1684.

Graham Neubig and Kevin Duh. 2013. How much is said in a tweet? a multilingual, information-theoretic perspective. In *AAAI Spring Symposium on Analyzing Microtext*, pages 32–39.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *HLT-NAACL*, pages 181–189.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *ACL-HLT*, pages 763–772.

Marc Schröder. 2001. Emotional speech synthesis: a review. In Paul Dalsgaard, Brge Lindberg, Henrik Benner, and Zheng-Hua Tan, editors, *INTERSPEECH*, pages 561–564.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proc. of International Conference on Spoken Language Processing*, pages 901–904.

Aobo Wang and Min-Yen Kan. 2013. Mining informal language from chinese microtext: Joint word recognition and segmentation. In *ACL*, pages 731–741.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *ACL-HLT*, pages 955–964.