Semi-Supervised Iterative Approach for Domain-Specific Complaint Detection in Social Media

Akash Gautam
MIDAS, IIIT-Delhi
akash15011@iiitd.ac.in

Debanjan Mahata
Bloomberg, New York, U.S.A
dmahata@bloomberg.net

Rakesh Gosangi
Bloomberg, New York, U.S.A
rgosangi@bloomberg.net

Rajiv Ratn Shah
MIDAS, IIIT-Delhi
rajivratn@iiitd.ac.in

Abstract

In this paper, we present a semi-supervised bootstrapping approach to detect product or service related complaints in social media. Our approach begins with a small collection of annotated samples which are used to identify a preliminary set of linguistic indicators pertinent to complaints. These indicators are then used to expand the dataset. The expanded dataset is again used to extract more indicators. This process is applied for several iterations until we can no longer find any new indicators. We evaluated this approach on a Twitter corpus specifically to detect complaints about transportation services. We started with an annotated set of 326 samples of transportation complaints, and after four iterations of the approach, we collected 2,840 indicators and over 3,700 tweets. We annotated a random sample of 700 tweets from the final dataset and observed that nearly half the samples were actual transportation complaints. Lastly, we also studied how different features based on semantics, orthographic properties, and sentiment contribute towards the prediction of complaints.

1 Introduction

Social media has lately become one of the primary venues where users express their opinions about various products and services. These opinions are extremely useful in understanding the user's perceptions and sentiment about these services. They are also useful in identifying potential defects (Abrahams et al., 2012) and thus critical to the execution of downstream customer service responses. Therefore, automatic detection of user complaints on social media could prove beneficial to both the clients and the service providers. To build such detection systems, we could employ supervised approaches that would typically require a large corpus of labeled training samples. However, labeling social media posts that capture complaints about a particular service is challenging because of their low prevalence and also the vast amounts of inevitable noise (Kietzmann et al., 2011; Lee, 2018). Additionally, social media platforms are also likely to be plagued with redundancy, where the posts are rephrased or structurally morphed before being re-posted (Ellison et al., 2011; Harrigan et al., 2012).

Prior work in event detection (Ritter et al., 2012) has demonstrated that simple linguistic indicators (phrases or n-grams) can be useful in the accurate discovery of events in social media. Though user complaints are not the same as events, more of a speech act (Preotiuc-Pietro et al., 2019), we posit that similar indicators can be used in complaint detection. To pursue this hypothesis, we propose a semi-supervised iterative approach to identify social media posts that complain about a specific service.

In our approach, we first begin with a small, manually curated dataset containing samples of social media posts complaining about a service. We then identify linguistic indicators (phrases or n-grams) that serve as strong evidence of this phenomenon. These indicators are then used to extract more posts from the unannotated corpus. This newly obtained data is then used to create a new set of indicators. This process is repeated until it reaches a certain convergence point. Since the set of indicators is growing after each iteration, they are re-evaluated continuously in terms of their relevance. This process is similar to the mutual bootstrapping approach for information extraction proposed in (Riloff et al., 2003).

We employ this approach to the problem of complaint detection for transportation services on Twitter. Transportation and its related logistic services are critical aspects of every economy as they account for nearly 40% of the value of international

Proceedings of the 3rd Workshop on e-Commerce and NLP (ECNLP 3), pages 46–53

trade (Rodrigue, 2007). As with most businesses (Gallaugher and Ransbotham, 2010; Gottipati et al., 2018), transportation also often relies on social media to ascertain feedback and initiate appropriate responses (Stelzer et al., 2016, 2014). In our experimental work, we started with an annotated set of 326 samples of transportation complaints, and after four iterations of the approach, we collected 2,840 indicators and over 3,700 tweets. We annotated a random sample of 700 tweets from the final dataset and observed that over 47% of the samples were actual transportation complaints. We also characterize the performance of basic classification algorithms on this dataset. In doing so, we also study how different linguistic features contribute to the performance of a supervised model in this domain.

The main contributions of this paper are as follows:

- We propose a semi-supervised iterative approach to collect user complaints about a service from social media platforms.
- We evaluate the proposed approach for the problem of complaint detection for transportation services on Twitter.
- We annotate a random sample of the resulting dataset to establish that nearly half the tweets were actual complaints.
- We release a curated dataset for the task of traffic-related complaint detection in social media[1].
- Lastly, we characterize the performance of basic classification algorithms on the dataset.

2 Related Work

Complaints are often considered dialogue acts used to express a mismatch between the expectation and reality (Olshtain and Weinbach, 1985). The problem of complaint detection is of great interest to the marketing and research teams of various service providers. Previous works on complaint identification have applied text mining with LDA and sentiment analysis on user-generated content (Liu et al., 2017; Duan et al., 2013). Prior works have also focused on leveraging data streamed from social media platforms for *outage* and complaint detection as they are publicly available (Augustine et al., 2012; Kursar and Gopinath, 2013).

(Yang et al., 2019) inspected customer support dialogue for support. Different complaint expressions have been explored by analyzing variations across cultures (Cohen and Olshtain, 1993), sociodemographic traits (Boxer, 1993) and temporal representations (Raghavan, 2014). However, mentioned works on user-generated content have focused on static data repositories only. These have not been robust to linguistic variations (Shah and Zimmermann, 2017) and morphological changes (Abdul-Mageed and Korayem, 2010). Our pipeline builds on linguistic identifiers to expand on lexical cues in order to identify complaint relevant posts.

Researches have proposed many semi-supervised architectures for identification of events pertaining to societal and civil unrest (Hua et al., 2013), using speech modality (Serizel et al., 2018; Wu et al., 2014; Zhang et al., 2017) and Hidden Markov Models (Zhang, 2005). These have been documented to give better performance as compared against their counterparts (Lee et al., 2017; Zheng et al., 2017) with minimal intervention (Rahimi et al., 2018). For our analysis, the semi-supervised approach has been preferred as opposed to supervised ones because: (a) usage of supervised approach relies on carefully choosing the training set making it cumbersome and less attractive for practical use (Watanabe, 2018) and (b) imbalance between the subjective and objective classes lead to poor performance (Yu et al., 2015).

3 Methods and Data

Our proposed approach begins with a large corpus of transport-related tweets and a small set of annotated complaints. We use this labeled data to create a set of *seed* indicators that drive the rest of our iterative complaint detection process.

3.1 Seed Data

We focused our experimentation over the period of November 2018 to December 2018. Our first step towards creating a corpus of transport-related tweets is to identify linguistic markers related to the transport domain. To this end, we scraped random posts from transport-related web forums[2]. These forums involve users discussing their grievances and raising awareness about a wide

[1]The dataset can be found at https://github.com/midas-research/transport-complaint-detection

[2]https://www.theverge.com/forums/transportation

array of transportation-related issues. We then processed this data to extract words and phrases (unigrams, bigrams, and trigrams) with high tf-idf scores. We then had human annotators prune them further to remove duplicates and irrelevant items. This resulted in a lexicon of 75 unique phrases. Some examples include *cabs, discount, tickets, underground, luggage, transit, parking, neighborhood, downtown, traffic, Uber.*

We used Twitter's public streaming API to query for tweets that contained any of the 75 phrases over the chosen time range. We then excluded non-English tweets and any tweets with less than two tokens. This resulted in a collection of 19,300 tweets. We will refer to this collection as corpus C. We chose a random sample of 1,500 tweets from this collection for human annotation. We employed two human annotators to identify traffic-related complaints from these 1,500 tweets. Following are some high-level details of the annotation task.

We instructed the annotators to identify any tweets that contain first-hand accounts of a complaint or a grievance related to a public/private mode of transport. Following is a sample tweet from this instruction: "*@[UserHandle] can you please make sure that compartment A-6 is at least clean before public use.*" We also instructed them to identify tweets that provide verifiable sources of information (news) about transport-related services. Sample tweet: "*4 hour jam in [place] area due to rain and poor management of traffic police.*". Lastly, we also explicitly asked them to exclude tweets that contain announcements or advertisements about transportation services. Sample tweet: "*Please use [name] cabs, you will get 60% discount on your first 3 rides.*"

The two annotators worked independently, and when we finally tallied their responses, we observed that they had an inter-annotator agreement rate of $\kappa = 0.81$ (Cohen kappa). In cases where the annotators disagreed, the labels were resolved through a discussion. After the disagreements were resolved, the final seed dataset had 326 samples of traffic-related complaints. We will refer to this as T_s. Table 1 shows some examples of tweets that were annotated as complaints.

3.2 Iterative Complaint Detection

Our proposed iterative approach is summarized in Algorithm 1. First, we use the seed data T_s to build a set of linguistic indicators I for complaints. We then use these indicators to get potential new complaints T_l from the corpus C. We merge T_s and T_l to build our new dataset. We then use this new dataset to extract a new set of indicators I_l. The indicators are combined with the original indicators I to extract the next version of T_l. This process is repeated until we can no longer find any new indicators.

Algorithm 1: Iterative Complaint Detection

Given: Corpus: C, Seed data: T_s
Get indicators I from T_s
$T = T_s$
Complaint Detection loop
Step 1: Select set T_l from C using I
Step 2: $T = T \cup T_l$
Step 3: Get indicators I_l from T
Step 4: $I = I \cup I_l$
Step 4: $C = C - T_l$

3.2.1 Extracting linguistic indicators

As shown in Algorithm 1, extracting linguistic indicators (n-grams) is one of the most important steps in the process. These indicators are critical to identifying tweets that are most likely domain-specific complaints. We employ two different approaches for extracting these indicators. For seed data, T_s, which is annotated, we just select n-grams with the highest tf-idf scores. In our experimental work, T_s had 326 annotated tweets. We identified 50 n-grams with the highest tf-idf scores to initialize I. Some examples include: ***problem, station, services, toll-fee, reply, fault, provide information, driver, district, passenger***. In subsequent iterations, when we are handling unannotated samples, we use a more advanced **domain relevance** criterion for extracting the indicators.

When extracting indicators from T_l, which is not annotated, it is possible that there could be frequently occurring phrases that are not necessarily indicative of complaints. These phrases could lead to a concept drift in subsequent iterations. To avoid these digressions, we use a measure of **domain relevance** when selecting indicators. This is defined as the ratio of the frequency of an n-gram in T_l to that of in T_r. T_r is a collection of randomly chosen tweets that do not intersect with C. In our experimental work, we defined T_r as a random sample of 5,000 tweets from a different time range than that of C. We also wanted to quantitatively en-

Samples of transport-related complaints.
1. *No metro fares will be reduced, but proper fare structure needs to be introduced .... right?.*
2. *It takes [name] govt. longer to refund charges, but it took them a few mins to remove that bus stop. You can't erase the problem[name].*
3. *I tried to lodge a complaint on [url] but see the results. Sir if 8 A.C's are not working in this coach , why have you attached that coach.*
4. *[name] Is that for when people can't travel due to your staff having to strike to keep everyone safe? Or perhaps short formed trains that you cant get on.*

Table 1: Sample tweets annotated as transport-related complaints.

sure that the lexicon in T_r is different from that of C. Namely, we calculated the cosine similarity between the two datasets in the tf-idf space. The cosine similarity at a value of 0.028 was statistically significant with a Pearson correlation coefficient value 0.012 (*p-value 0.0034*) (Schober et al., 2018).

3.2.2 Selection of tweets

Given a set of indicators I, the process of selecting tweets from corpus C is fairly straightforward. It only requires to identify all the tweet that contains any of the indicators. The only caveat here is to reduce the redundancy in the dataset. For this, we just filtered out tweets that have a cosine similarity of more than 0.85 with any other tweet in the tf-idf space (Albakour et al., 2013). This process also helped remove tweets, which are exact matches, sub-strings, or differing by some punctuation. Removal of these redundant tweets also helps in diversifying the lexicon for subsequent iterations.

3.2.3 Complaints dataset

Our iterative approach converged in four rounds, after which it did not extract any new indicators. Figure 1 shows the counts of indicators and the number of tweets after each iteration. After four iterations, this approach chose 3,732 tweets and generated 2,840 unique indicators. We also manually inspected the indicators chosen during the process. We observed that only indicators with a domain relevance score of ≥ 2.5 were chosen for subsequent iterations. Table 2 provides a few examples of strong and weak indicators acquired after the first iteration. In this figure, strong indicators are those with a domain relevance score ≥ 2.5.

We chose a random set of 700 tweets from the final complaints dataset T and annotated them manually to help understand the quality. We used the same guidelines as discussed in section 3.1 and also employed the same annotators as before. The annotators once again obtained a high agreement score of $\kappa = 0.83$. After resolving the disagreements, we observed that 332 tweets were labeled as complaints. This accounts for 47.4% of the sampled 700 tweets. This demonstrates that nearly half the tweets selected by our semi-supervised approach were traffic-related complaints. This is a significantly higher proportion in the original seed data T_s, where only 21.7% were actual complaints.

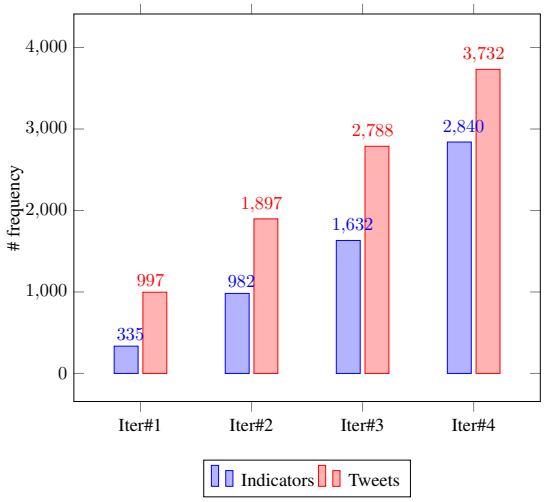

Figure 1: Number of indicators and tweets collected after each iteration.

4 Modeling

We conducted a series of experiments to understand if we can automatically build simple machine learning models to detect complaints. These experiments also helped us evaluate the quality of the final dataset. Additionally, this experimental work also studies how different types of linguistic features contribute to the detection of social media complaints. For these experiments, we used the annotated sample of 700 posts as a test dataset. We built our training dataset by selecting another 2,000 posts from the original corpus C, and anno-

Strength	Indicator
Strong	car travel (*5.80*), your complaint (*3.62*), technical problem (*3.59*), report officer (*3.44*), traffic control (*3.33*), make apologies(*3.29*)
Weak	you go (*0.55*), sure (*0.51*), please (*0.49*), take this (*0.44*), with you (*0.42*), therefore (*0.39*), make him (*0.36*)

Table 2: Examples of some strong and weak indicators. The numbers in brackets denote the respective domain relevance score.

Feature	Accuracy(%)	F1-score
Semantic Features		
Unigrams	**75.3**	**0.70**
POS Tags	70.1	0.66
Word2Vec cluster	72.1	0.67
Pronoun Types	69.6	0.65
Sentiment Features		
MPQA	68.2	0.61
NRC	67.9	0.59
VADER	68.0	0.62
Stanford Sentiment	68.7	0.63
Orthographic Features		
Textual Meta-data	69.3	0.62
Intensifiers	72.5	0.67
Request Features		
Request Model	70.1	0.66
Politeness Markers	70.4	0.63

Table 3: Predictive accuracy and F1-score associated with different types of features.

tated them once again per guidelines discussed in section 3.1. In this sample, we observed that the annotators had similar agreements scores of $\kappa = 0.79$, and there were 702 instances of complaints.

4.1 Features

We also wanted to understand the predictive power of different types of linguistic features towards the detection of complaints. These features can be broadly broken down into four groups. (i) The first group of features are based on simple semantic properties such as *n-grams*, *word embeddings*, and *part of speech tags*. (ii) The second group of features are based on *pre-trained sentiment models* or *lexicons*. (iii) The third group of features use *orthographic information* such as *hashtags*, *user mentions*, and *intensifiers*. (iv) The last group of features again use *pre-trained models* or *lexicons* associated with *request*, which is a closely related speech act (Švárová, 2008).

4.1.1 Semantic features

We experimented with four different semantic features:

Unigrams: Each tweet (Wallach, 2006) is represented as sparse vector of tf-idf values corresponding to the constituent tokens.

Word2Vec Clusters: We follow the same approach as in (Preoţiuc-Pietro et al., 2015), where words are clustered using pair-wise similarities in Word2Vec space (Mikolov et al., 2013). Each tweet is then represented as a distribution over these clusters; the values are proportional to the number of tokens belonging to a cluster. These clusters have previously been demonstrated to have great interpretability (Preoţiuc-Pietro et al., 2015, 2017; Zou et al., 2016).

POS Tags: We used the Stanford POS Tagger (Manning et al., 2014) to represent tweets as a dense frequency vector over five main POS tags: *nouns, adjectives, adverbs, verbs, pronouns*.

Pronoun Types: Pronouns are often used in complaints and suggestions to reveal personal involvement or to add intensity to an opinion (Claridge, 2007; Meinl, 2013). We identify various pronoun types (*first person, second person, third person, demonstrative, indefinite*) using dictionaries and use their counts as features.

4.1.2 Sentiment features

We expect sentiment to contribute strongly towards the prediction of complaints. We experiment with two *pre-trained models*: Stanford Sentiment (Socher et al., 2013) and VADER (Hutto and Gilbert, 2014). Namely, we use the scores predicted by these models as representations of tweets. Likewise, we also experiment with two sentiment lexicons: MPQA (Wilson et al., 2005), NRC (Mohammad et al., 2013) for assigning sentiment scores to tweets.

4.1.3 Orthographic features

Our first set of orthographic feature uses counts of URLs, hashtags, user mentions, and special symbols used in the post. The second set of orthographic features try to identify potential intensifiers such as capitalization and repeated use of exclamation or question marks. These types of intensifiers are often used to express anger or strong opinions

(Meinl, 2013).

4.1.4 Request features

A request is a speech act very closely related to complaints. Often, the main motivation behind a complaint on a social media platform is to get a correction or reparation from the service providers (Blum-Kulka and Olshtain, 1984). We use the model presented in (Danescu-Niculescu-Mizil et al., 2013) to detect if a given tweet is a *request*. Requests might also often include polite phrases in expectation of better service. They are coded using various dictionaries e.g, downgraders (*little*), down-toners (*just*), hedges (*somewhat*). Apology markers have the same effect as politeness markers, they may include greetings at the start (*Good Morning*), direct start (*e.g so*), subjunctive phrases (*could you*) (Švárová, 2008). We utilize pre-defined dictionaries to determine the presence of politeness identifiers along with the politeness score of the tweet based on the model in (Danescu-Niculescu-Mizil et al., 2013).

4.2 Results

We trained a logistic regression model for complaint detection using each one of the features described in section 4.1. Table 3 summarizes the results in terms of accuracy and macro averaged F1-score. The best performing model is based on unigrams, with an accuracy of 75.3%. There is not a significant difference in the performance of different sentiment models. It is also interesting to observe that simple features like the counts of different pronoun types and counts of intensifiers have strong predictive ability. Overall, we observe that most of the features studied here have some ability to predict complaints.

5 Conclusion and Future Work

In this paper, we presented a semi-supervised iterative approach for the detection of complaints in social media platforms. The process begins with a small sample of annotated examples, and then iteratively builds more linguistic identifiers to expand the dataset. We evaluated this approach on the domain of transportation on Twitter, starting with a sample of 326 annotated tweets. After four iterations, we were able to construct a corpus with over 3,700 tweets. Annotation of random samples established that nearly half the tweets were actual complaints. We evaluated the predictive power based on semantic, orthographic, and sentiment features. We observed that complaint is a complex speech act, which is related to many other linguistic properties.

Automatic detection of complaints is not only useful to service providers as feedback; it could also prove helpful in improving service providers' operations and in downstream applications such as developing chat-bots. Additionally, it could also be of interest to linguists in understanding how humans express grievances and criticism.

This proposed methodology could be applied to many other products or services to detect complaints. This would only additionally require some lexicons and a small annotated dataset. We also expect it would be fairly straightforward to adapt this technique to many other types of speech acts. Further investigation is necessary to understand how this method compares against supervised or completely unsupervised techniques.

References

Muhammad Abdul-Mageed and Mohammed Korayem. 2010. Automatic identification of subjectivity in morphologically rich languages: the case of arabic. *Computational approaches to subjectivity and sentiment analysis*, 2:2–6.

Alan S Abrahams, Jian Jiao, G Alan Wang, and Weiguo Fan. 2012. Vehicle defect discovery from social media. *Decision Support Systems*, 54(1):87–97.

M Albakour, Craig Macdonald, Iadh Ounis, et al. 2013. On sparsity and drift for effective real-time filtering in microblogs. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 419–428. ACM.

Eriq Augustine, Cailin Cushing, Alex Dekhtyar, Kevin McEntee, Kimberly Paterson, and Matt Tognetti. 2012. Outage detection via real-time social stream analysis: leveraging the power of online complaints. In *Proceedings of the 21st International Conference on World Wide Web*, pages 13–22.

Shoshana Blum-Kulka and Elite Olshtain. 1984. Requests and apologies: A cross-cultural study of speech act realization patterns (ccsarp). *Applied linguistics*, 5(3):196–213.

Diana Boxer. 1993. Social distance and speech behavior: The case of indirect complaints. *Journal of pragmatics*, 19(2):103–125.

Claudia Claridge. 2007. The superlative in spoken english. In *Corpus Linguistics 25 Years on*, pages 121–148. Brill Rodopi.

Andrew D Cohen and Elite Olshtain. 1993. The production of speech acts by efl learners. *Tesol Quarterly*, 27(1):33–56.

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 *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. ACM.

Wenjing Duan, Qing Cao, Yang Yu, and Stuart Levy. 2013. Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In *2013 46th Hawaii International Conference on System Sciences*, pages 3119–3128. IEEE.

Nicole B Ellison, Jessica Vitak, Charles Steinfield, Rebecca Gray, and Cliff Lampe. 2011. Negotiating privacy concerns and social capital needs in a social media environment. In *Privacy online*, pages 19–32. Springer.

John Gallaugher and Sam Ransbotham. 2010. Social media and customer dialog management at starbucks. *MIS Quarterly Executive*, 9(4).

Swapna Gottipati, Venky Shankararaman, and Jeff Rongsheng Lin. 2018. Text analytics approach to extract course improvement suggestions from students' feedback. *Research and Practice in Technology Enhanced Learning*, 13(1):6.

Nicholas Harrigan, Palakorn Achananuparp, and Ee-Peng Lim. 2012. Influentials, novelty, and social contagion: The viral power of average friends, close communities, and old news. *Social Networks*, 34(4):470–480.

Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. 2013. Sted: semi-supervised targeted-interest event detectionin in twitter. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1466–1469.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Jan H Kietzmann, Kristopher Hermkens, Ian P McCarthy, and Bruno S Silvestre. 2011. Social media? get serious! understanding the functional building blocks of social media. *Business horizons*, 54(3):241–251.

Brian Kursar and Jayadev Gopinath. 2013. Validating customer complaints based on social media postings. US Patent App. 13/646,548.

In Lee. 2018. Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons*, 61(2):199–210.

Kathy Lee, Ashequl Qadir, Sadid A Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. 2017. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 705–714.

Xia Liu, Alvin C Burns, and Yingjian Hou. 2017. An investigation of brand-related user-generated content on twitter. *Journal of Advertising*, 46(2):236–247.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Marja E Meinl. 2013. *Electronic complaints: an empirical study on British English and German complaints on eBay*, volume 18. Frank & Timme GmbH.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Elite Olshtain and Liora Weinbach. 1985. *Complaints-A study of speech act behavior among native and nonnative speakers of Hebrew*. Tel Aviv University.

Daniel Preotiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. *arXiv preprint arXiv:1906.03890*.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. 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)*, volume 1, pages 1754–1764.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740.

Preethi Raghavan. 2014. *Medical event timeline generation from clinical narratives*. Ph.D. thesis, The Ohio State University.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised user geolocation via graph convolutional networks. *arXiv preprint arXiv:1804.08049*.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 25–32.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.

Jean-Paul Rodrigue. 2007. Transportation and globalization. *Encyclopedia of*.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, and Ankit Parag Shah. 2018. Large-scale weakly labeled semi-supervised sound event detection in domestic environments. *arXiv preprint arXiv:1807.10501*.

Rajiv Shah and Roger Zimmermann. 2017. *Multimodal analysis of user-generated multimedia content*. Springer.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Anselmo Stelzer, Frank Englert, Stephan Hörold, and Cindy Mayas. 2014. Using customer feedback in public transportation systems. In *2014 International Conference on Advanced Logistics and Transport (ICALT)*, pages 29–34. IEEE.

Anselmo Stelzer, Frank Englert, Stephan Hörold, and Cindy Mayas. 2016. Improving service quality in public transportation systems using automated customer feedback. *Transportation Research Part E: Logistics and Transportation Review*, 89:259–271.

Jana Švárová. 2008. *Politeness markers in spoken language*. Ph.D. thesis, Masarykova univerzita, Pedagogická fakulta.

Hanna M Wallach. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM.

Kohei Watanabe. 2018. Newsmap: A semi-supervised approach to geographical news classification. *Digital Journalism*, 6(3):294–309.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. 2014. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2665–2672.

Wei Yang, Luchen Tan, Chunwei Lu, Anqi Cui, Han Li, Xi Chen, Kun Xiong, Muzi Wang, Ming Li, Jian Pei, et al. 2019. Detecting customer complaint escalation with recurrent neural networks and manually-engineered features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 56–63.

Z. Yu, R. K. Wong, C. Chi, and F. Chen. 2015. A semi-supervised learning approach for microblog sentiment classification. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 339–344.

Dingwen Zhang, Junwei Han, Lu Jiang, Senmao Ye, and Xiaojun Chang. 2017. Revealing event saliency in unconstrained video collection. *IEEE Transactions on Image Processing*, 26(4):1746–1758.

Zhu Zhang. 2005. Mining inter-entity semantic relations using improved transductive learning. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Xin Zheng, Aixin Sun, Sibo Wang, and Jialong Han. 2017. Semi-supervised event-related tweet identification with dynamic keyword generation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1619–1628.

Bin Zou, Vasileios Lampos, Russell Gorton, and Ingemar J Cox. 2016. On infectious intestinal disease surveillance using social media content. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 157–161. ACM.