Automatically Identifying Complaints in Social Media

Daniel Preoţiuc-Pietro
Bloomberg LP
dpreotiucpie@bloomberg.net

Mihaela Găman
Politehnica University of Bucharest
mpgaman@gmail.com

Nikolaos Aletras
University of Sheffield
n.aletras@sheffield.ac.uk

Abstract

Complaining is a basic speech act regularly used in human and computer mediated communication to express a negative mismatch between reality and expectations in a particular situation. Automatically identifying complaints in social media is of utmost importance for organizations or brands to improve the customer experience or in developing dialogue systems for handling and responding to complaints. In this paper, we introduce the first systematic analysis of complaints in computational linguistics. We collect a new annotated data set of written complaints expressed in English on Twitter.[1] We present an extensive linguistic analysis of complaining as a speech act in social media and train strong feature-based and neural models of complaints across nine domains achieving a predictive performance of up to 79 F1 using distant supervision.

1 Introduction

Complaining is a basic speech act used to express a negative mismatch between reality and expectations towards a state of affairs, product, organization or event (Olshtain and Weinbach, 1987). Understanding the expression of complaints in natural language and automatically identifying them is of utmost importance for: (a) linguists to obtain a better understanding of the context, intent and types of complaints on a large scale; (b) psychologists to identify human traits underpinning complaint behavior and expression; (c) organizations and advisers to improve the customer service by identifying and addressing client concerns and issues effectively in real time, especially on social media; (d) developing downstream natural language processing (NLP) applications, such as dialogue systems that aim to automatically identify complaints.

Tweet	C	S
@FC_Help hi, I ordered a necklace over a week ago and it still hasn't arrived (...)	✓	
@BootsUK I love Boots! Shame you're introducing a man tax of 7% in 2018 :(	✓	✓
You suck		✓

Table 1: Examples of tweets annotated for complaint (C) and sentiment (S).

However, complaining has yet to be studied using computational approaches. The speech act of complaining, as previously defined in linguistics research (Olshtain and Weinbach, 1987) and adopted in this study, has as its core the concept of violated or breached expectations i.e., the person posting the complaint had their favorable expectations breached by a party, usually the one to which the complaint is addressed.

Complaints have been previously analyzed by linguists (Vásquez, 2011) as distinctly different from expressing negative sentiment towards an entity. Key to the definition of complaints is the expression of the breach of expectations. Table 1 shows examples of tweets highlighting the differences between complaints and sentiment. The first example expresses the writer's breach of expectations about an item that was expected to arrive, but does not express negative sentiment toward the entity, while the second shows mixed sentiment and expresses a complaint about a tax that was introduced. The third statement is an insult that implies negative sentiment, but there are not enough cues to indicate any breach of expectations; hence, this cannot be categorized as a complaint.

This paper presents the first extensive analysis of complaints in computational linguistics. Our contributions include:

1. The first publicly available data set of complaints extracted from Twitter with expert annotations spanning nine domains (e.g., software,

[1]Data and code is available here: https://github.com/danielpreotiuc/complaints-social-media

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5008–5019
Florence, Italy, July 28 - August 2, 2019.

transport);
2. An extensive quantitative analysis of the syntactic, stylistic and semantic linguistic features distinctive of complaints;
3. Predictive models using a broad range of features and machine learning models, which achieve high predictive performance for identifying complaints in tweets of up to 79 F1;
4. A distant supervision approach to collect data combined with domain adaptation to boost predictive performance.

2 Related Work

Complaints have to date received significant attention in linguistics and marketing research. Olshtain and Weinbach (1987) provide one of the early definitions of a complaint as when a speaker expects a favorable event to occur or an unfavorable event to be prevented and these expectations are breached. Thus, the discrepancy between the expectations of the complainer and the reality is the key component of identifying complaints.

Complaining is considered to be a distinct speech act, as defined by speech act theory (Austin, 1975; Searle, 1969) which is central to the field of pragmatics. Complaints are either addressed to the party responsible for enabling the breach of expectations (direct complaints) or indirectly mention the party (indirect complaints) (Boxer, 1993b). Complaints are widely considered to be among the face-threatening acts (Brown and Levinson, 1987) – acts that aim to damage the face or self-esteem of the person or entity the act is directed at. The concept of face (Goffman, 1967) represents the public image specific of each person or entity and has two aspects: positive (i.e., the desire to be liked) and negative face (i.e., the desire to not be imposed upon). Complaints can intrinsically threaten both positive and negative face. Positive face of the responsible party is affected by having enabled the breach of expectations. Usually, when a direct complaint is made, the illocutionary function of the complaint is to request for a correction or reparation for these events. Thus, this aims to affect negative face by aiming to impose an action to be undertaken by the responsible party. Complaints usually co-occur with other speech acts such as warnings, threats, suggestions or advice (Olshtain and Weinbach, 1987; Cohen and Olshtain, 1993).

Previous linguistics research has qualitatively examined the types of complaints elicited via discourse completion tests (DCT) (Trosborg, 1995) and in naturally occurring speech (Laforest, 2002). Differences in complaint strategies and expression were studied across cultures (Cohen and Olshtain, 1993) and socio-demographic traits (Boxer, 1993a). In naturally occurring text, the discourse structure of complaints has been studied in letters to editors (Hartford and Mahboob, 2004; Ranosa-Madrunio, 2004). In the area of linguistic studies on computer mediated communication, Vásquez (2011) performed an analysis of 100 negative reviews on TripAdvisor, which showed that complaints in this medium often co-occur with other speech acts including positive and negative remarks, frequently make explicit references to expectations not being met and directly demand a reparation or compensation. Meinl (2013) studied complaints in eBay reviews by annotating 200 reviews in English and German with the speech act sequence that makes up each complaint e.g., warning, annoyance (the annotations are not available publicly or after contacting the authors). Mikolov et al. (2018) analyze which financial complaints submitted to the Consumer Financial Protection Bureau will receive a timely response. Most recently, Yang et al. (2019) studied customer support dialogues and predicted if these complaints will be escalated with a government agency or made public on social media.

To the best of our knowledge, the only previous work that tackles a concept defined as a complaint with computational methods is by Zhou and Ganesan (2016) which studies Yelp reviews. However, they define a complaint as a 'sentence with negative connotation with supplemental information'. This definition is not aligned with previous research in linguistics (as presented above) and represents only a minor variation on sentiment analysis. They introduce a data set of complaints, unavailable at the time of this submission, and only perform a qualitative analysis, without building predictive models for identifying complaints.

3 Data

To date, there is no available data set with annotated complaints as previously defined in linguistics (Olshtain and Weinbach, 1987). Thus, we create a new data set of written utterances annotated with whether they express a complaint. We use Twitter as the data source because (1) it represents

a platform with high levels of self-expression; and (2) users directly interact with other users or corporate brand accounts. Tweets are openly available and represent a popular option for data selection in other related tasks such as predicting sentiment (Rosenthal et al., 2017), affect (Mohammad et al., 2018), emotion analysis (Mohammad and Kiritchenko, 2015), sarcasm (González-Ibánez et al., 2011; Bamman and Smith, 2015), stance (Mohammad et al., 2016), text-image relationship (Vempala and Preoţiuc-Pietro, 2019) or irony (Van Hee et al., 2016; Cervone et al., 2017; Van Hee et al., 2018).

3.1 Collection

We choose to manually annotate tweets in order to provide a solid benchmark to foster future research on this task.

Complaints represent a minority of the total written posts on Twitter. We use a data sampling method that increases the hit rate of complaints, following previous work on labeling infrequent linguistic phenomena such as irony (Mohammad et al., 2018). Numerous companies use Twitter to provide customer service and address user complaints. We select tweets directed to these accounts as candidates for complaint annotation. We manually assembled a list of 93 customer service handles. Using the Twitter API,[2] we collected all the tweets that are available to download (the most recent 3,200). We then identified all the original tweets to which the customer support handle responded. We randomly sample an equal number of tweets addressed to each customer support handle for annotation. Using this method, we collected 1,971 tweets to which the customer support handles responded.

Further, we have also manually grouped the customer support handles in several high-level domains based on their industry type and area of activity. We have done this to enable analyzing complaints by domain and assess transferability of classifiers across domains. In related work on sentiment analysis, reviews for products from four different domains were collected across domains in a similar fashion (Blitzer et al., 2007). All customer support handles grouped by category are presented in Table 2.

We add to our data set randomly sampled tweets to ensure that there is a more representative and diverse set of tweets for feature analysis and to ensure that the evaluation does not disproportionally contain complaints. We thus additionally sampled 1,478 tweets consisting of two groups of 739 tweets: the first group contains random tweets addressed to any other Twitter handle (at-replies) to match the initial sample, while the second group contains tweets not addressed to a Twitter handle.

As preprocessing, we anonymize all usernames present in the tweet and URLs and replace them with placeholder tokens. To extract the unigrams used as features, we use DLATK, which handles social media content and markup such as emoticons or hashtags (Schwartz et al., 2017). Tweets were filtered for English using langid.py (Lui and Baldwin, 2012) and retweets were excluded.

3.2 Annotation

We create a binary annotation task for identifying if a tweet contains a complaint or not. Tweets are short and usually express a single thought. Therefore, we consider the entire tweet as a complaint if it contains at least one complaint speech act. For annotation, we adopt as the guideline a complaint definition similar to that from previous linguistic research (Olshtain and Weinbach, 1987; Cohen and Olshtain, 1993): *"A complaint presents a state of affairs which breaches the writer's favorable expectation"*.

Each tweet was labeled by two independent annotators, authors of the paper, with significant experience in linguistic annotation. After an initial calibration run of 100 tweets (later discarded from the final data set), each annotator labeled all 1,971 tweets independently. The two annotators achieved a Cohen's Kappa $\kappa = 0.731$, which is in the upper part of the *substantial* agreement band (Artstein and Poesio, 2008). Disagreements were discussed and resolved between the annotators. In total, 1,232 tweets (62.4%) are complaints and 739 are not complaints (37.6%). The statistics for each category is in Table 3.

4 Features

In our analysis and predictive experiments, we use the following groups of features: generic linguistic features proven to perform well in text classification tasks (Preoţiuc-Pietro et al., 2015; Preoţiuc-Pietro et al., 2017; Volkova and Bell, 2017; Preoţiuc-Pietro and Ungar, 2018) (unigrams, LIWC, word clusters), methods for predict-

[2] https://developer.twitter.com/

Food & Beverage	Apparel	Retail	Cars	Services	Software & Online Services	Transport	Electronics	Other
ABCustomerCare	NeimanMarcus	HarrodsService	HondaCustSvc	GEICO_Service	YelpSupport	AirAsiaSupport	AskPlayStation	BlackandDecker
ArbysCares	FC_Help	BN_Care	VWCares	Safaricom_Care	UbisoftSupport	SEPTA_Social	XBoxSupport	WhirlpoolCare
KFC_UKI_Help	Zara_Care	WalmartHelp	ChryslerCares	VirginMedia	SqSupportUK	FreaterAnglia	LenovoSupport	NYTCare
McDonalds	NBaStoreSupport	BootsHelp	SubaruCustCare	ThreeUKSupport	AWSSupport	RailMinIndia	AppleSupport	WashPostHelp
PizzaHut	HM_CustServ	WholeFoods	AlfaRomeoCares	KenyaPower_Care	SHO_Help	VirginTrains	Moto_Support	MACCosmetics
	SupportAtTommy	BestBuySupport		GeorgiaPower	TeamTurboTax	Delta	OnePlus_Support	HolidayInn
	BurberyService	IKEAUSSupport		UPShelp	DropboxSupport	British_Airways	SamsungSupport	
	Nordstrom	AmazonHelp		ComcastCares	AdobeCare	JetBlue	FitbitSupport	
	DSGsupport	AskEBay		AOLSupportHelp	Uber_Support	United	BeatsSupport	
	TopmanAskUs			EE	NortonSupport	AmericanAir	NvidiaCC	
	SuperDry_Care			VodafoneIN	MediumSupport	SouthwestAir	HPSupport	
	ASOS_HereToHelp			BTcare	TwitterSupport		NikeSupport	
				HMRCCustomers	Hulu_Support			
				DirecTVService	MicrosoftHelps			

Table 2: List of customer support handles by domain. The domain is chosen based on the most frequent product or service the account usually receives complaints about (e.g., NikeSupport receives most complaints about the Nike Fitness Bands).

Category	Complaints	Not Complaints
Food & Beverage	95	35
Apparel	141	117
Retail	124	75
Cars	67	25
Services	207	130
Software & Online Services	189	103
Transport	139	109
Electronics	174	112
Other	96	33
Total	1232	739

Table 3: Number of tweets annotated as complaints across the nine domains.

ing sentiment or emotion which have an overlap with complaints and complaint specific features which capture linguistic aspects typical of complaints (Meinl, 2013; Danescu-Niculescu-Mizil et al., 2013):

Unigrams. We use the bag-of-words approach to represent each tweet as a TF-IDF weighted distribution over the vocabulary consisting of all words present in at least two tweets (2,641 words).

LIWC. Traditional psychology studies use dictionary-based approaches to representing text. The most popular method is based on Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) consisting of 73 manually constructed lists of words (Pennebaker et al., 2015) including parts-of-speech, topical or stylistic categories. Each tweet is thus represented as a distribution over these categories.

Word2Vec Clusters. An alternative to LIWC for identifying semantic themes in a tweet is to use automatically generated word clusters. These clusters can be thought of as *topics* i.e., groups of words that are semantically and/or syntactically similar. The clusters help reduce the feature space and provide good interpretability (Lampos et al., 2014; Preoţiuc-Pietro et al., 2015; Preoţiuc-Pietro et al., 2015; Lampos et al., 2016; Aletras and Chamberlain, 2018). We follow Preoţiuc-Pietro et al. (2015) to compute clusters using spectral clustering (Shi and Malik, 2000) applied to a word-word similarity matrix weighted with the cosine similarity of the corresponding word embedding vectors (Mikolov et al., 2013). The clusters help reduce the feature space and provide good interpretability.[3] For brevity and clarity, we present experiments using 200 clusters as in (Preoţiuc-Pietro et al., 2015). We aggregated all the words in a tweet and represent each tweet as a distribution of the fraction of tokens belonging to each cluster.

Part-of-Speech Tags. We analyze part-of-speech tag usage to quantify the syntactic patterns associated with complaints and to enhance the representation of unigrams. We part-of-speech tag all tweets using the Twitter model of the Stanford Tagger (Derczynski et al., 2013). In prediction experiments we supplement each unigram feature with their POS tag (e.g., *I_PRP*, *bought_VBN*). For feature analysis, we represent each tweet as a bag-of-words distribution over part-of-speech unigrams and bigrams in order to uncover regular syntactic patterns specific of complaints.

Sentiment & Emotion Models. We use existing sentiment and emotion analysis models to study their relationship to complaint annotations and to measure their predictive power on our complaint data set. If the concepts of negative sentiment and complaint were to coincide, standard sentiment prediction models that have access to larger sets of training data should be very competitive on predicting complaints. We test the following models:

- **MPQA**: We use the MPQA sentiment lexicon (Wiebe et al., 2005) to assign a positive and negative score to each tweet based on the ratio of tokens in a tweet which appear in the positive and negative MPQA lists respectively. These scores are used as features.
- **NRC**: We use the word lexicon derived using

[3] We have tried other alternatives to building clusters: using NPMI (Bouma, 2009), GloVe (Pennington et al., 2014) and LDA (Blei et al., 2003).

crowd-sourcing from (Mohammad and Turney, 2010, 2013) for assigning to each tweet the proportion of tokens that have positive, negative and neutral sentiment, as well as one of eight emotions that include the six basic emotions of Ekman (Ekman, 1992) (anger, disgust, fear, joy, sadness and surprise) plus trust and anticipation. All scores are used as features in prediction in order to maximize their predictive power.

- **Volkova & Bachrach (V&B)**: We quantify positive, negative and neutral sentiment as well as the six Ekman emotions for each message using the model made available in (Volkova and Bachrach, 2016) and use them as features in predicting complaints. The sentiment model is trained on a data set of 19,555 tweets that combine all previously annotated tweets across seven public data sets.
- **VADER**: We use the outcome of the rule-based sentiment analysis model which has shown very good predictive performance on predicting sentiment in tweets (Gilbert and Hutto, 2014).
- **Stanford**: We quantify sentiment using the Stanford sentiment prediction model as described in (Socher et al., 2013).

Complaint Specific Features. The features in this category are inspired by linguistic aspects specific to complaints (Meinl, 2013):

- **Request.** The illocutionary function of complaints is often that of requesting for a correction or reparation for the event that caused the breach of expectations (Olshtain and Weinbach, 1987). We explicitly predict if an utterance is a request using the model introduced in (Danescu-Niculescu-Mizil et al., 2013).

- **Intensifiers.** In order to increase the face-threatening effect a complaint has on the complainee, intensifiers are usually used by the person expressing the complaint (Meinl, 2013). We use features derived from: (1) capitalization patterns often used online as an equivalent to shouting (e.g., number/percentage of capitalized words, number/percentage of words starting with capitals, number/percentage of capitalized letters); and (2) repetitions of exclamation marks, question marks or letters within the same token.

- **Downgraders and Politeness Markers.** In contrast to intensifiers, downgrading modifiers are used to reduce the face-threat involved when voicing a complaint, usually as part of a strategy to obtain a reparation for the breach of expectation (Meinl, 2013). Downgraders are coded by several dictionaries: play down (e.g., *i wondered if*), understaters (e.g., *one little*), disarmers (e.g., *but*), downtoners (e.g., *just*) and hedges (e.g., *somewhat*). Politeness markers have a similar effect to downgraders and include apologies (e.g., *sorry*), greetings at the start, direct questions, direct start (e.g., *so*), indicative modals (e.g., *can you*), subjunctive modals (e.g., *could you*), politeness markers (e.g., *please*) (Svarova, 2008) and politeness maxims (e.g., *i must say*). Finally, we directly predict the politeness score of the tweet using the model presented in (Danescu-Niculescu-Mizil et al., 2013).

- **Temporal References.** Temporal references are often used in complaints to stress how long a complainer has been waiting for a correction or reparation from the addressee or to provide context for their complaint (e.g., mentioning the date in which they have bought an item) (Meinl, 2013). We identify time expressions in tweets using SynTime, which achieved state-of-the-art results across on several benchmark data sets (Zhong et al., 2017). We represent temporal expressions both as days elapsed relative to the day of the post and in buckets of different granularities (one day, week, month, year).

- **Pronoun Types.** Pronouns are used in complaints to reveal the personal involvement or opinion of the complainer and intensify or reduce the face-threat of the complaint based on the person or type of the pronoun (Claridge, 2007; Meinl, 2013). We split pronouns using dictionaries into: first person, second person, third person, demonstrative (e.g., *this*) and indefinite (e.g., *everybody*).

5 Linguistic Feature Analysis

This section presents a quantitative analysis of the linguistic features distinctive of tweets containing complains in order to gain linguistic insight into this task and data. We perform analysis of all previously described feature sets using univariate Pearson correlation (Schwartz et al., 2013). We compute correlations independently for each feature between its distribution across messages (features are first normalized to sum up to unit for each message) and a variable encoding if the tweet was annotated as a complaint or not.

Top unigrams and part-of-speech features specific of complaints and non-complaints are presented in Table 4. The top features for the LIWC

Complaints		Not Complaints	
Feature	r	**Feature**	r
Unigrams			
not	.154	<URL>	.150
my	.131	!	.082
working	.124	he	.069
still	.123	thank	.067
on	.119	,	.064
can't	.113	love	.064
service	.112	lol	.061
customer	.109	you	.060
why	.108	great	.058
website	.107	win	.058
no	.104	'	.058
?	.098	she	.054
fix	.093	:	.053
won't	.092	that	.053
been	.090	more	.052
issue	.089	it	.052
days	.088	would	.051
error	.087	him	.047
is	.084	life	.046
charged	.083	good	.046
POS (Unigrams and Bigrams)			
VBN	.141	UH	.104
$	.118	NNP	.098
VBZ	.114	PRP	.076
NN_VBZ	.114	HT	.076
PRP$	.107	PRP_.	.076
PRP$_NN	.105	PRP_RB	.067
VBG	.093	NNP_NNP	.062
CD	.092	VBP_PRP	.054
WRB_VBZ	.084	JJ	.053
VBZ_VBN	.084	DT_JJ	.051

Table 4: Features associated with complaint and non-complaint tweets, sorted by Pearson correlation (r) computed between the normalized frequency of each feature and the complaint label across all tweets. All correlations are significant at $p < .01$, two-tailed t-test, Simes corrected.

categories and Word2Vec topics are presented in Table 5. All correlations shown in these tables are statistically significant at $p < .01$, with Simes correction for multiple comparisons.

Negations. Negations are uncovered through unigrams (*not*, *no*, *won't*) and the top LIWC category (*NEGATE*). Central to complaining is the concept of breached expectations. Hence the complainers use negations to express this discrepancy and to describe their experience with the product or service that caused this.

Issues. Several unigrams (*error*, *issue*, *working*, *fix*) and a cluster (*Issues*) contain words referring to issues or errors. However, words regularly describing negative sentiment or emotions are not one of the most distinctive features for complaints. On the other hand, the presence of terms that show positive sentiment or emotions (*good*, *great*, *win*, *POSEMO*, *AFFECT*, *ASSENT*) are among the top most distinctive features for a tweet not being labeled as a complaint. In addition, other words and clusters expressing positive states such as gratitude (*thank*, *great*, *love*) or laughter (*lol*) are also distinctive for tweets that are not complaints.

Linguistics research on complaints in longer documents identified that complaints are likely to co-occur with other speech acts, including with expressions of positive or negative emotions (Vásquez, 2011). In our data set, perhaps due to the particular nature of Twitter communication and the character limit, complainers are much more likely to not express positive sentiment in a complaint and do not regularly post negative sentiment. Instead, they choose to focus more on describing the issue regarding the service or product in an attempt to have it resolved.

Pronouns. Across unigrams, part-of-speech patterns and word clusters, we see a distinctive pattern emerging around pronoun usage. Complaints use more possessive pronouns, indicating that the user is describing personal experiences. A distinctive part-of-speech pattern common in complaints is possessive pronouns followed by nouns (*PRP$_NN*) which refer to items of services possessed by the complainer (e.g., *my account*, *my order*). Complaints tend to not contain personal pronouns (*he*, *she*, *it*, *him*, *you*, *SHEHE*, *MALE*, *FEMALE*), as the focus on expressing the complaint is on the self and the party the complaint is addressed to and not other third parties.

Punctuation. Question marks are distinctive of complaints, as many complaints are formulated as questions to the responsible party (e.g., *why is this not working?*, *when will I get my response?*). Complaints are not usually accompanied by exclamation marks. Although exclamation marks are regularly used for emphasis in the context of complaints, most complainers in our data set prefer not to use them perhaps in an attempt to address them in a less confrontational manner.

Temporal References. Mentions of time are specific of complaints (*been*, *still*, *on*, *days*, *Temporal References* cluster). Their presence is usually needed to provide context for the event that caused the breach of expectations. Another role of temporal references is to express dissatisfaction towards non-responsiveness of the responsible party in addressing their previous requests. In addition, the presence of verbs in past participle (*VBN*) is the most distinctive part-of-speech pattern of complaints. These are used to describe actions com-

Complaints			Not Complaints		
Label	**Words**	r	**Label**	**Words**	r
LIWC Features					
NEGATE	not, no, can't, don't, never, nothing, doesn't, won't	.271	POSEMO	thanks, love, thank, good, great, support, lol, win	.185
RELATIV	in, on, when, at, out, still, now, up, back, new	.225	AFFECT	thanks, love, thank, good, great, support, lol	.111
FUNCTION	the, i, to, a, my, and, you, for, is, in	.204	SHEHE	he, his, she, her, him, he's, himself	.105
TIME	when, still, now, back, new, never, after, then, waiting	.186	MALE	he, his, man, him, sir, he's, son	.086
DIFFER	not, but, if, or, can't, really, than, other, haven't	.169	FEMALE	she, her, girl, mom, ma, lady, mother, female, mrs	.084
COGPROC	not, but, how, if, all, why, or, any, need	.132	ASSENT	yes, ok, awesome, okay, yeah, cool, absolutely, agree	.080
Word2Vec Clusters					
Cust. Service	service, customer, contact, job, staff, assist, agent	.136	Gratitude	thanks, thank, good, great, support, everyone, huge, proud	.089
Order	order, store, buy, free, delivery, available, package	.128	Family	old, friend, family, mom, wife, husband, younger	.063
Issues	delayed, closed, between, outage, delay, road, accident	.122	Voting	favorite, part, stars, model, vote, models, represent	.060
Time Ref.	been, yet, haven't, long, happened, yesterday, took	.122	Contests	Christmas, gift, receive, entered, giveaway, enter, cards	.058
Tech Parts	battery, laptop, screen, warranty, desktop, printer	.100	Pets	dogs, cat, dog, pet, shepherd, fluffy, treats	.054
Access	use, using, error, password, access, automatically, reset	.098	Christian	god, shall, heaven, spirit, lord, belongs, soul, believers	.053

Table 5: Group text features associated with tweets that are complaints and not complaints. Features are sorted by Pearson correlation (r) between their each feature's normalized frequency and the outcome. We restrict to only the top six categories for each feature type. All correlations are significant at $p < .01$, two-tailed t-test, Simes corrected. Within each cluster, words are sorted by frequency in our data set. Labels for Word2Vec clusters are assigned by the authors.

pleted in the past (e.g., *i've bought*, *have come*) in order to provide context for the complaint.

Verbs. Several part-of-speech patterns distinctive of complaints involve present verbs in third person singular (*VBZ*). In general, these verbs are used in complaints to reference an action that the author expects to happen, but his expectations are breached (e.g., *nobody is answering*). Verbs in gerund or present participle are used as a complaint strategy to describe things that just happened to a user (e.g., *got an email saying my service will be terminated*).

Topics. General topics typical of complaint tweets include requiring assistance or customer support. Several groups of words are much more likely to appear in a complaint, although not used to express complaints per se: about orders or deliveries (in the retail domain), about access (in complaints to service providers) and about parts of tech products (in tech). This is natural, as people are more likely to deliberately tweet about an order or tech parts if they want to complain about them. This is similar to sentiment analysis, where not only emotionally valenced words are predictive of sentiment.

6 Predicting Complaints

In this section, we experiment with different approaches to build predictive models of complaints from text content alone. We first experiment with feature based approaches including Logistic Regression classification with Elastic Net regularization (LR) (Zou and Hastie, 2005).[4] We train the classifiers with all individual feature types.

Neural Methods. For reference, we experiment with two neural architectures. In both architectures, tweets are represented as sequences of one-hot word vectors which are first mapped into embeddings. A multi-layer perceptron (MLP) network (Hornik et al., 1989) feeds the embedded representation ($E = 200$) of the tweet (mean embedding of its constituent words) into a dense hidden layer ($D = 100$) followed by a ReLU activation function and dropout (0.2). The output layer is one dimensional dense layer with a sigmoid activation function. The second architecture, a Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network, processes sequentially the tweet by modeling one word (embedding) at each time step followed by the same output layer as in MLP. The size of the hidden state of the LSTM is $L = 50$. We train the networks using the Adam optimizer (Kingma and Ba, 2014) (learning rate is set to 0.01) by minimizing the binary cross-entropy.

Experimental Setup. We conduct experiments using a nested stratified 10-fold cross-validation, where nine folds are used for training and one for testing (i.e., outer loop). In the inner loop, we choose the model parameters[5] using a 3-fold cross-validation on the tweets from the nine folds of training data (from the outer loop). Train/dev/test splits for each experiment are released together with the data for replicability. We report predictive performance of the models as the mean accuracy, F1 (macro-averaged) and ROC AUC over the 10 folds (Dietterich, 1998).

[4] We use the Scikit Learn implementation (Pedregosa et al., 2011).

[5] We tune the regularization term, α and the mixing parameter of the LR model. For the neural networks, we tune the size of the embedding E, the dense hidden layer D, the LSTM cells L and the learning rate of the optimizer.

Model	Acc	F1	AUC
Most Frequent Class	64.2	39.1	0.500
Logistic Regression			
Sentiment – MPQA	64.2	39.1	0.499
Sentiment – NRC	63.9	42.2	0.599
Sentiment – V&B	68.9	60.0	0.696
Sentiment – VADER	66.0	54.2	0.654
Sentiment – Stanford	68.0	55.6	0.696
Complaint Specific (all)	65.7	55.2	0.634
Request	64.2	39.1	0.583
Intensifiers	64.5	47.3	0.639
Downgraders	65.4	49.8	0.615
Temporal References	64.2	43.7	0.535
Pronoun Types	64.1	39.1	0.545
POS Bigrams	72.2	66.8	0.756
LIWC	71.6	65.8	0.784
Word2Vec Clusters	67.7	58.3	0.738
Bag-of-Words	79.8	77.5	0.866
All Features	**80.5**	**78.0**	**0.873**
Neural Networks			
MLP	78.3	76.2	0.845
LSTM	80.2	77.0	0.864

Table 6: Complaint prediction results using **logistic regression** (with different types of linguistic features), **neural network** approaches and the **most frequent class** baseline. Best results are in bold.

Results. Results are presented in Table 6. Most sentiment analysis models show accuracy above chance in predicting complaints. The best results are obtained by the Volkova & Bachrach model (Sentiment – V&B) which achieves 60 F1. However, models trained using linguistic features on the training data obtain significantly higher predictive accuracy. Complaint specific features are predictive of complaints, but to a smaller extent than sentiment, reaching an overall 55.2 F1. From this group of features, the most predictive groups are intensifiers and downgraders. Syntactic part-of-speech features alone obtain higher performance than any sentiment or complaint feature group, showing the syntactic patterns discussed in the previous section hold high predictive accuracy for the task. The topical features such as the LIWC dictionaries (which combine syntactic and semantic information) and Word2Vec topics perform in the same range as the part of speech tags. However, best predictive performance is obtained using bag-of-word features, reaching an F1 of up to 77.5 and AUC of 0.866. Further, combining all features boosts predictive accuracy to 78 F1 and 0.864 AUC. We notice that neural network approaches are comparable, but do not outperform the best performing feature-based model, likely in part due to the training data size.

Model	Acc	F1	AUC
Most Frequent Class	64.2	39.1	0.500
LR-All Features – Original Data	80.5	78.0	0.873
Dist. Supervision + Pooling	77.2	75.7	0.853
Dist. Supervision + EasyAdapt	**81.2**	**79.0**	**0.885**

Table 7: Complaint prediction results using the original data set and distantly supervised data. All models are based on logistic regression with bag-of-word and Part-of-Speech tag features.

Distant Supervision. We explore the idea of identifying extra complaint data using distant supervision to further boost predictive performance. Previous work has demonstrated improvements on related tasks relying on weak supervision e.g., in the form of tweets with related hashtags (Bamman and Smith, 2015; Volkova and Bachrach, 2016; Cliche, 2017). Following the same procedure, seven hashtags were identified with the help of the training data to likely correspond to complaints: #appallingcustomercare, #badbusiness, #badcustomerserivice, #badservice, #lostbusiness, #unhappycustomer, #worstbrand. Tweets were collected to contain these hashtags from a combination of the 1% Twitter archive between 2012-2018 and by filtering tweets with these hashtags in real-time from Twitter REST API for three months. We collected in total 18,218 tweets (excluding retweets and duplicates) equated to complaints. As negative complaint examples, the same amount of tweets were sampled randomly from the same time interval. All hashtags were removed and the data was preprocessed identically as the annotated data set.

We experiment with two techniques for combining distantly supervised data with our annotated data. First, the tweets obtained through distant supervision are simply added to the annotated training data in each fold (**Pooling**). Secondly, as important signal may be washed out if the features are joined across both domains, we experiment with domain adaptation using the popular EasyAdapt algorithm (Daumé III, 2007) (**EasyAdapt**). Experiments use logistic regression with bag-of-word features enhanced with part-of-speech tags, because these performed best in the previous experiment.

Results presented in Table 7 show that the domain adaptation approach further boosts F1 by 1 point to 79 (t-test, $p<0.5$) and ROC AUC by 0.012. However, simply pooling the data actually hurts predictive performance leading to a drop of more than 2 points in F1.

Domain	In-Domain	Pooling	EasyAdapt
Food & Beverage	63.9	60.9	**83.1**
Apparel	**76.2**	71.1	72.5
Retail	58.8	**79.7**	**79.7**
Cars	41.5	77.8	**80.9**
Services	65.2	75.9	**76.7**
Software	61.3	73.4	**78.7**
Transport	56.4	**73.4**	69.8
Electronics	66.2	73.0	**76.2**
Other	42.4	**82.8**	**82.8**

Table 8: Performance of models in Macro F1 on tweets from each domain.

Test / Train	F&B	A	R	Ca	Se	So	T	E	O
Food & Bev.	–	58.1	52.5	66.4	59.7	58.9	54.1	61.4	53.7
Apparel	63.9	–	74.4	65.1	70.8	71.2	68.5	76.9	85.6
Retail	58.8	74.4	–	70.1	72.6	69.9	68.7	69.6	82.7
Cars	68.7	61.1	65.1	–	58.8	67.	59.3	62.9	68.2
Services	65.	74.2	75.8	74.	–	68.8	74.2	77.9	77.9
Software	62.	74.2	68.	67.9	72.8	–	72.8	72.1	80.6
Transport	59.3	71.7	72.4	67.	74.6	75.	–	72.6	81.7
Electronics	61.6	75.2	71.	68.	75.	69.9	68.2	–	78.7
Other	56.1	71.3	72.4	70.2	73.5	67.2	68.5	71.	–
All	70.3	77.7	79.5	82.0	79.6	80.1	76.8	81.7	88.2

Table 9: Performance of models trained with tweets from one domain and tested on other domains. All results are reported in ROC AUC. The **All** line displays results on training on all categories except the category in testing.

Domain Experiments We assess the performance of models trained using the best method and features by using in training: (1) using only in-domain data (**In-Domain**); (2) adding out-of-domain data into the training set (**Pooling**); and (3) combining in- and out-of-domain data with EasyAdapt domain adaptation (**EasyAdapt**). The experimental setup is identical to the one described in the previous experiments. Table 8 shows the model performance in macro-averaged F1 using the best performing feature set.

Results show that, in all but one case, adding out-of-domain data helps predictive performance. The apparel domain is qualitatively very different from the others as a large number of complaints are about returns or the company not stocking items, hence leading to different features being important for prediction. Domain adaptation is beneficial the majority of domains, lowering performance on a single domain compared to data pooling. This highlights the differences in expressing complaints across domains. Overall, predictive performance is high across all domains, with the exception of transport.

Cross Domain Experiments

Finally, Table 9 presents the results of models trained on tweets from one domain and tested on all tweets from other domains, with additional models trained on tweets from all domains except the one that the model is tested on.

We observe that predictive performance is relatively consistent across all domains with two exceptions ('Food & Beverage' consistently shows lower performance, while 'Other' achieves higher performance) when using all the data available from the other domains.

7 Conclusions & Future Work

We presented the first computational approach using methods from computational linguistics and machine learning to modeling complaints as defined in prior studies in linguistics and pragmatics (Olshtain and Weinbach, 1987). To this end, we introduced the first data set consisting of English Twitter posts annotated with complaints across nine domains. We analyzed the syntactic patterns and linguistic markers specific of complaints. Then, we built predictive models of complaints in tweets using a wide range of features reaching up to 79% Macro F1 (0.885 AUC) and conducted experiments using distant supervision and domain adaptation to boost predictive performance. We studied performance of complaint prediction models on each individual domain and presented results with a domain adaptation approach which overall improves predictive accuracy. All data and code is available to the research community to foster further research on complaints.

A predictive model for identification of complaints is useful to companies that wish to automatically gather and analyze complaints about a particular event or product. This would allow them to improve efficiency in customer service or to more cheaply gauge popular opinion in a timely manner in order to identify common issues around a product launch or policy proposal.

In the future, we plan to identify the target of the complaint in a similar way to aspect-based sentiment analysis (Pontiki et al., 2016). We plan to use additional context and conversational structure to improve performance and identify the socio-demographic covariates of expressing and phrasing complaints. Another research direction is to study the role of complaints in personal conversation or in the political domain, e.g., predicting political stance in elections (Tsakalidis et al., 2018).

Acknowledgments

Nikolaos Aletras is supported by an Amazon AWS Cloud Credits for Research award.

References

Nikolaos Aletras and Benjamin Paul Chamberlain. 2018. Predicting Twitter User Socioeconomic Attributes with Network and Language Information. In *Proceedings of the 29th on Hypertext and Social Media*, HT, pages 20–24.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

John Langshaw Austin. 1975. *How to do Things with Words*. Oxford University Press.

David Bamman and Noah A Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *Proceedings of the 9th International Conference on Weblogs and Social Media*, ICWSM, pages 574–577.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-Boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL, pages 440–447.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Diana Boxer. 1993a. Complaints as Positive Strategies: What the Learner Needs to Know. *Tesol Quarterly*, 27(2):277–299.

Diana Boxer. 1993b. Social Distance and Speech Behavior: The Case of Indirect Complaints. *Journal of Pragmatics*, 19(2):103–125.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge University Press.

Alessandra Cervone, Evgeny A Stepanov, Fabio Celli, and Giuseppe Riccardi. 2017. Irony detection: from the twittersphere to the news space. In *CLiC-it 2017-Italian Conference on Computational Linguistics*, volume 2006.

Claudia Claridge. 2007. Constructing a Corpus from the Web: Message Boards. *Language and Computers*, 59(87).

Mathieu Cliche. 2017. BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2017)*, *SEM, pages 573–580.

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, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL, pages 250–259.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 256–263.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP, pages 198–206.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3-4):169–200.

CJ Gilbert and Eric Hutto. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International Conference on Weblogs and Social Media*, ICWSM, pages 216–225.

Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Interaction*. Aldine.

Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying Sarcasm in Twitter: A Closer Look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, ACL, pages 581–586.

Beverly Hartford and Ahmar Mahboob. 2004. Models of Discourse in the Letter of Complaint. *World Englishes*, 23(4):585–600.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Marty Laforest. 2002. Scenes of Family Life: Complaining in Everyday Conversation. *Journal of Pragmatics*, 34(10-11):1595–1620.

Vasileios Lampos, Nikolaos Aletras, Jens K. Geyti, Bin Zou, and Ingemar J. Cox. 2016. Inferring the Socioeconomic Status of Social Media Users Based on Behaviour and Language. In *Advances in Information Retrieval*, pages 689–695.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and Characterising User Impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 405–413.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 system demonstrations*, ACL, pages 25–30.

Marja E Meinl. 2013. *Electronic Complaints: An Empirical Study on British English and German Complaints on eBay*, volume 18. Frank & Timme GmbH.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 746–751.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2018. Deconfounded Lexicon Induction for Interpretable Social Science. In *Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 746–751.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, *SEM, pages 31–41.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, *SEM, pages 1–17.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, NAACL, pages 26–34.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Elite Olshtain and Liora Weinbach. 1987. Complaints: A Study of Speech Act Behavior among Native and Non-native Speakers of Hebrew. *Bertuccelli-Papi, M. (Eds.), The Pragmatic Perspective: Selected Papers from the 1985 International Pragmatics Conference*, pages 195–208.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *JMLR*, 12.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Austin, TX: Pennebaker Conglomerates.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Mahway: Lawrence Erlbaum Associates.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1532–1543.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. SemEval-2016 Task 5: Aspect based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

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*, ACL, pages 1754–1764.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-Level Race and Ethnicity Predictors from Twitter Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING, pages 1534–1545.

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*, ACL, pages 729–740.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.

Marilu Ranosa-Madrunio. 2004. The Discourse Organization of Letters of Complaint to Editors in Philippine English and Singapore English. *Philippine Journal of Linguistics*, 35(2):67–97.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, *SEM, pages 502–518.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin EP Seligman. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS ONE*, 8(9).

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Johannes Eichstaedt, and Lyle Ungar. 2017. DLATK: Differential Language Analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP, pages 55–60.

John R Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*, volume 626. Cambridge University Press.

Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

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*, EMNLP, pages 1631–1642.

Jana Svarova. 2008. *Politeness Markers in Spoken Language*. Ph.D. thesis, Masarykova Univerzita.

Anna Trosborg. 1995. *Interlanguage Pragmatics: Requests, Complaints, and Apologies*, volume 7. Walter de Gruyter.

Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. 2018. Nowcasting the stance of social media users in a sudden vote: The case of the Greek Referendum. CIKM, pages 367–376.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016. Monday Mornings are my Fave:)# not Exploring the Automatic Recognition of Irony in English tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2730–2739.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. Semeval-2018 Task 3: Irony detection in English Tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, *SEM, pages 39–50.

Camilla Vásquez. 2011. Complaints Online: The case of TripAdvisor. *Journal of Pragmatics*, 43(6):1707–1717.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring Perceived Demographics from User Emotional Tone and User-Environment Emotional Contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 1567–1578.

Svitlana Volkova and Eric Bell. 2017. Identifying Effective Signals to Predict Deleted and Suspended Accounts on Twitter across Languages. ICWSM, pages 290–298.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.

Wei Yang, Luchen Tan, Chunwei Lu, Anqi Cui, Han Li, Xi Chen, Kun Xiong, Muzi Wang, Ming Li, Jian Pei, and Jimmy Lin. 2019. Detecting Customer Complaint Escalation with Recurrent Neural Networks and Manually-Engineered Features. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Industry Track)*, NAACL, pages 56–63.

Xiaoshi Zhong, Aixin Sun, and Erik Cambria. 2017. Time expression analysis and recognition using syntactic token types and general heuristic rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 420–429.

Guangyu Zhou and Kavita Ganesan. 2016. Linguistic Understanding of Complaints and Praises in User Reviews. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, NAACL, pages 109–114.

Hui Zou and Trevor Hastie. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.