Linguistically Informed Tweet Categorization for Online Reputation Management

Gerard Lynch and Pádraig Cunningham
Centre for Applied Data Analytics Research
(CeADAR)
University College Dublin
Belfield Office Park
Dublin 4, Ireland
firstname.lastname@ucd.ie

Abstract

Determining relevant content automatically is a challenging task for any aggregation system. In the business intelligence domain, particularly in the application area of Online Reputation Management, it may be desirable to label tweets as either customer comments which deserve rapid attention or tweets from industry experts or sources regarding the higher-level operations of a particular entity. We present an approach using a combination of linguistic and Twitter-specific features to represent tweets and examine the efficacy of these in distinguishing between tweets which have been labelled using Amazon's Mechanical Turk crowdsourcing platform. Features such as part-of-speech tags and function words prove highly effective at discriminating between the two categories of tweet related to several distinct entity types, with Twitter-related metrics such as the presence of hashtags, retweets and user mentions also adding to classification accuracy. Accuracy of 86% is reported using an SVM classifier and a mixed set of the aforementioned features on a corpus of tweets related to seven business entities.

1 Motivation

Online Reputation Management (ORM) is a growing field of interest in the domain of business intelligence. Companies and individuals alike are highly interested in monitoring the opinions of others across social and traditional media and this information can have considerable business value for corporate entities in particular.

1.1 Challenges

There are a number of challenges in creating an end-to-end software solution for such purposes, and several shared tasks have already been established to tackle these issues[1]. The most recent RepLab evaluation was concerned with four tasks related to ORM, *filtering*, *polarity for reputation*, *topic detection* and *priority assignment*. Based on these evaluations, it is clear that although the state of the art of topic-based filtering of tweets is relatively accomplished (Perez-Tellez et al., 2011; Yerva et al., 2011; Spina et al., 2013), other aspects of the task such as sentiment analysis and prioritisation of tweets based on content are less trivial and require further analysis.

Whether Twitter mentions of entities are actual customer comments or in fact represent the views of traditional media or industry experts and sources is an important distinction for ORM systems. With this study we investigate the degree to which this task can be automated using supervised learning methods.

2 Related Work

2.1 Studies on Twitter data

While the majority of research in the computational sciences on Twitter data has focused on issues such as topic detection (Cataldi et al., 2010), event detection, (Weng and Lee, 2011; Sakaki et al., 2010), sentiment analysis, (Kouloumpis et al., 2011), and other tasks based primarily on the topical and/or semantic content of tweets, there is a growing body of work which investigates more subtle forms of information represented in tweets, such as reputation and trustworthiness, (O'Donovan et al., 2012), authorship attribution (Layton et al., 2010; Bhargava et al., 2013) and Twitter spam detection, (Benevenuto et al., 2010).

[1] See (Amigó et al., 2012) and (Amigó et al., 2013) for details of the RepLab series

Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 73–78,
Baltimore, Maryland, USA. June 27, 2014.

These studies combine Twitter-specific and textual features such as retweet counts, tweet lengths and hashtag frequency, together with sentence-length, character n-grams and punctuation counts.

2.2 Studies on non-Twitter data

The textual features used in our work such as n-grams of words and parts-of-speech have been used for gender-based language classification (Koppel et al., 2002), social profiling and personality type detection (Mairesse et al., 2007), native language detection from L2 text, (Brooke and Hirst, 2012) translation source language detection, (van Halteren, 2008; Lynch and Vogel, 2012) and translation quality detection, (Vogel et al., 2013).

3 Experimental setup and corpus

Tweets were gathered between June 2013 and January 2014 using the *twitter4j* Java library. A language detector was used to filter only English-language tweets.[2] The criteria for inclusion were that the entity name was present in the tweet. The entities focused on in this study had relatively unambigious business names, so no complex filtering was necessary.

3.1 Pilot study

A smaller pilot study was carried out before the main study in order to examine response quality and accuracy of instruction. Two hundred sample tweets concerning two airlines[3] were annotated using Amazon's Mechanical Turk system by fourteen Master annotators. After annotation, we selected the subset (72%) of tweets for which both annotators agreed on the category to train the classifier. During the pilot study, the tweets were pre-processed[4] to remove @ and # symbols and punctuation to treat account names and hashtags as words. Hyperlinks representations were maintained within the tweets. The Twitter-specific metrics were not employed in the pilot study.

3.2 Full study

In the full study, 2454 tweets concerning seven business entities[5] were tagged by forty annotators as to whether they corresponded to one of the three categories described in Section 1.1. For 57% of the tweets, annotators agreed on the categories with disagreement in the remaining 43%. The disputed tweets were annotated again by two annotators. From this batch, a similar proportion were agreed on. For the non-agreed tweets in the second round, a majority category vote was reached by combining the four annotations over the first and second rounds. After this process, roughly two hundred tweets remained as ambiguous (each having two annotations for one of two particular categories) and these were removed from the corpus used in the experiments.

3.3 Category breakdown

Table 5 displays the number of tweets for which no majority category agreement was reached. The majority disagreement class across all entities are texts which have been labelled as both business operations and other. For the airline entities, a large proportion of tweets were annotated as both customer comment and other, this appeared to be a categorical issue which may have required clarification in the instructions. The smallest category for tied agreement is customer comment and business operations, it appears that the distinction between these categories was clearer based on the data provided to annotators. 2078 tweets were used in the final experiments. The classes were somewhat imbalanced for the final corpus, the *business operations* category was the largest, with 1184 examples, *customer comments* contained 585 examples and the *other* category contained 309 examples.

3.4 Feature types

The features used for classification purposes can be divided into the following two categories:

1. **Twitter-specific**:
   - Tweet is a retweet or not
   - Tweet contains a mention
   - Tweet contains a hashtag or a link
   - Weight measure (See Fig 3)
   - Retweet account for a tweet.
2. **Linguistic**: The linguistic features are based on the textual content of the tweet represented as word unigrams, word bigrams and part-of-speech bigrams.

[2] A small amount of non-English tweets were found in the dataset, these were assigned to the *Other category*.

[3] Aer Lingus and Ryanair

[4] This was not done in the full study, these symbols were counted and used as features.

[5] Aer Lingus, Ryanair, Bank of Ireland, C & C Group, Permanent TSB, Glanbia, Greencore

We used TagHelperTools, (Rosé et al., 2008) for textual feature creation which utilises the Stanford NLP toolkit for NLP annotation and returns formatted representations of textual features which can be employed in the Weka toolkit which implements various machine learning algorithms. All linguistic feature frequencies were binarised in our representations[6].

4 Results

4.1 Pilot study

Using the Naive Bayes classifier in the Weka toolkit and a feature set consisting of 130 word tokens, 80% classification accuracy was obtained using ten-fold cross validation on the full set of tweets . Table 1 shows the top word features when ranked using 10-fold cross validation and the information gain metric for classification power over the three classes. Using the top 50 ranked POS-bigram features alone, 74% classification accuracy was obtained using the Naive Bayes classifier. Table 2 shows the top twenty features, again ranked by information gain.

Combining the fifty POS-bigrams and the 130 word features, we obtained 84% classification accuracy using the Naive Bayes classifier. Accuracy was improved by removing all noun features from the dataset and using the top seventy five features from the remaining set ranked with information gain, resulting in 86.6% accuracy using the SVM classifier with a linear kernel. Table 3 displays the top twenty combined features.

Rank	Feature	Rank	Feature
1	http	11	investors
2	flight	12	would
3	talks	13	by
4	for	14	says
5	strike	15	profit
6	an	16	cabin
7	you	17	crew
8	I	18	via
9	that	19	at
10	action	20	since

Table 1: Top 20 ranked word features for pilot study

[6] 1 if feature is present in a tweet, otherwise 0.

Rank	Feature	Rank	Feature
1	NNP_EOL	11	VB_PRP
2	VBD_JJ	12	NN_NNS
3	NNP_VBD	13	IN_PRP$
4	NNP_NN	14	BOL_CD
5	BOL_PRP	15	BOL_JJS
6	VBD_NNP	16	IN_VBN
7	NNP_CC	17	PRP$_JJ
8	TO_NNP	18	PRP_MD
9	NN_RB	19	PRP$_VBG
10	RB_JJ	20	CC_VBP

Table 2: Top 20 ranked POS bigram features for pilot study

Rank	Feature	Rank	Feature
1	http	11	TO_NNP
2	NNP_EOL	12	RB_JJ
3	NNP_VBD	13	that
4	VBD_JJ	14	tells
5	NNP_NN	15	way
6	BOL_PRP	16	I
7	VBD_NNP	17	would
8	NNP_CC	18	you
9	for	19	NN_RB
10	an	20	BOL_JJS

Table 3: Top 20 ranked combined features for pilot study

4.2 Full study

4.2.1 Results

Using the SMO classifier, Weka's support vector machine implementation using a linear kernel, a hybrid feature set containing linguistic, custom and Twitter-specific features obtained 72% classification accuracy for the three categories. F-measures were highest for the *business operations* class, and lowest for the *other* class, which contained the most diversity. Examining Figure 2, it is clear that f-measures for the *other* class are almost zero. This indicates that tweets given this category may not be homogeneous enough to categorise using the features defined in Table 7.

4.3 Two classes

After the removal of the *other* class from the experiment, the same feature set obtained 86% classification accuracy between the two remaining classes. The distinguishing features consisted predominantly of pronouns (*I, me, my*), part-of-

Entity	BO	CC	Other
Aer Lingus	174	138	44
Ryanair	58	212	52
AIB	69	29	43
BOI	208	85	40
C&C	45	14	15
Glanbia	276	39	46
Greencore	37	4	13
Kerry Group	158	10	36
Permanent TSB	160	54	20

Table 4: Tweets per entity by category: Majority agreement

Entity	CC+BO	O-CC	O-BO
Aer Lingus	4	24	15
Ryanair	7	30	8
AIB	4	5	11
BOI	9	5	16
C&C	0	1	3
Glanbia	7	4	19
Greencore	0	0	2
Kerry Group	5	2	12
Permanent TSB	3	6	10

Table 5: Tweets per entity by category: Tied agreement

speech bigrams including pairs of plural nouns, lines beginning with prepositions and function words (*so, just, new, it*). Business operations tweets were more likely to mention a user account or be a retweet, personal pronouns were more commonplace in customer comments and as observed in the pilot study, customer comments were more likely to begin with a preposition and business operations tweets were more likely to contain noun-noun compounds and pairs of coordinating conjunctions and nouns.

4.4 Features

Hashtags were slightly more common in business operations tweets, however the number of hashtags was not counted, simply whether at least one was present. Hashtags as a proportion of words might be a useful feature for further studies. Function words and POS tags were highly discriminatory, indicating that this classifier may be applicable to different topic areas. Weight (See Figure 3) was a distinguishing feature, with business operations tweets having higher weight scores, reflecting the tendency for these tweets to originate from Twitter accounts linked to news sources or influential industry experts.

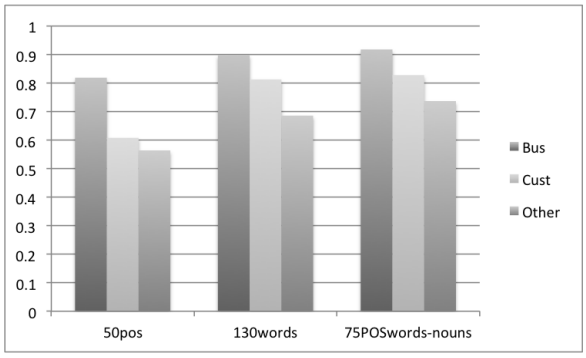

Figure 1: F-scores by category for pilot study

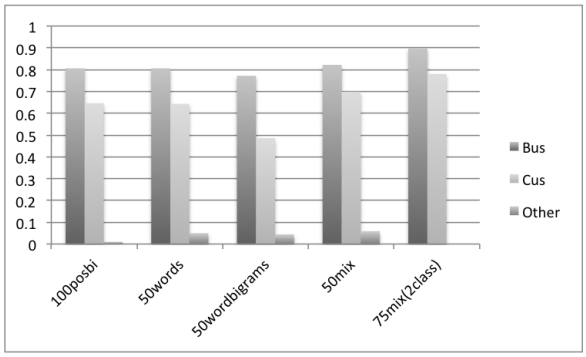

Figure 2: F-scores by category for full study

5 Results per sub-category

To investigate whether the entity domain had a bearing on the results, we separated the data into three subsets, airlines, banks and food industry concerns. We performed the same feature selection as in previous experiments, calculating each feature type separately, removing proper nouns, hashtags and account names from the word n-grams, then combining and ranking the features using ten-fold cross validation and information gain. The SVM classifier reported similar results to the main study on the three class problem for each sub-domain, and for the two class problem results ranged between 86-87% accuracy, similar

$$\frac{\text{Number of followers}}{\text{Number following}}(\text{retweets})$$

Figure 3: Twitter weight metric

to the results on the mixed set[7]. Thus, we believe that the individual subdomains do not warrant different classifiers for the problem, indeed examining the top 20-ranked features for each subdomain, there is a large degree of overlap, as seen in bold and italics in Table 6.

Banks	Airlines	Food
@	@	@
my	*NNP_NNP*	**PRP_VBP**
i	**i**	**i**
me	**BOL_IN**	**BOL_IN**
PRP_VBP	**PRP_VBP**	VB_PRP
account	*DT_NN*	BOL_PRP
NNP_VBZ	IN_PRP	HASHASH
VB_PRP	the	**you**
IN_PRP	new	**me**
you	PRP_VBD	know
BOL_RB	NNP_VBZ	my
RB_JJ	IN_DT	i_know
NNP_NNP	**you**	PRP_CC
PRP_VBD	BOL_PRP	used
my_bank	*ISRT*	BOL_CC
DT_NN	it	NNP_CD
NN_PRP	**me**	NN_NNP
VBD_PRP	my	CC_PRP
BOL_IN	RB_RB	*ISRT*
i'm	so	CC_NNP

Table 6: Top twenty ranked features by Information Gain for three domains

6 Conclusions and future directions

6.1 Classification results

We found that accurate categorization of our predefined tweet types was possible using shallow linguistic features. This was aided by Twitter specific metrics but these did not add significantly to the classification accuracy[8]. The lower score (72-73%) in the three class categorization problem is due to the linguistic diversity of the *other* tweet category.

6.2 Annotation and Mechanical Turk

We found the definition of categorization criteria to be an important and challenging step when using Mechanical Turk for annotation. The high degree of annotator disagreement reflected this, however it is important to note that in many cases, tweets fit equally into two or more of our defined categories. The use of extra annotations[9] allowed for agreement to be reached in the majority of cases, however employing more evaluations could have also resulted in deadlock. Examples of ambiguous tweets included: *Cheap marketing tactics. Well, if it ain't broke, why fix it! RT @Ryanair's summer '14 schedule is now on sale!* where a Twitter user has retweeted an official announcement and added their own comment.

Another possible pitfall is that as Mechanical Turk is a US-based service and requires workers to have a US bank account in order to perform work, Turkers tend to be US-based, and therefore an annotation task concerning non-US business entities is perhaps more difficult without sufficient background awareness of the entities in question.

Future experiments will apply the methodology developed here to a larger dataset of tweets, one candidate would be the dataset used in the RepLab 2013 evaluation series which contains 2,200 annotated tweets for 61 business entities in four domains.

Acknowledgments

The authors are grateful to Enterprise Ireland and the IDA for funding this research and CeADAR through their Technology Centre Programme.

Rank	Feature	Rank	Feature
1	@	26	NNP_PRP
2	i	27	NN_PRP
3	PRP_VBP	28	VBP_PRP
4	my	29	when
5	BOL_IN	30	if
6	me	31	don't
7	you	32	PRP_MD
8	NNP_NNP	33	they
9	IN_PRP	34	like
10	VB_PRP	35	PRP_VB
11	PRP_VBD	36	got
12	**WEIGHT**	37	CC_NNP
13	so	38	but
14	NNP_VBZ	39	RB_IN
15	BOL_PRP	40	**RT**
16	RB_JJ	41	with
17	DT_NN	42	PRP_IN
18	BOL_RB	43	a
19	it	44	NNS_RB
20	PRP_RB	45	CC_PRP
21	RB_RB	46	VBD_PRP
22	IN_DT	47	VBD_DT
23	i'm	48	no
24	just	49	the
25	get	50	PRP$_NN

Table 7: Top 50 ranked mixed features for main study

[7]The food subset was highly imbalanced however, containing only 43 customer comments and 313 business operations tweets, the other two subsets were relatively balanced.

[8]ca. 2% decrease in accuracy on removal.

[9]over the initial two annotators

References

Enrique Amigó, Adolfo Corujo, Julio Gonzalo, Edgar Meij, and Maarten de Rijke. 2012. Overview of replab 2012: Evaluating online reputation management systems. In *CLEF (Online Working Notes/Labs/Workshop)*.

Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2013. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 333–352. Springer.

Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgılio Almeida. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6.

Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. 2013. Stylometric analysis for authorship attribution on twitter. In *Big Data Analytics*, pages 37–47. Springer International Publishing.

Julian Brooke and Graeme Hirst. 2012. Measuring interlanguage: Native language identification with l1-influence metrics. In *LREC*, pages 779–784.

Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *ICWSM*.

Robert Layton, Paul Watters, and Richard Dazeley. 2010. Authorship attribution for twitter in 140 characters or less. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*, pages 1–8. IEEE.

Gerard Lynch and Carl Vogel. 2012. Towards the automatic detection of the source language of a literary translation. In *COLING (Posters)*, pages 775–784.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.(JAIR)*, 30:457–500.

John O'Donovan, Byungkyu Kang, Greg Meyer, Tobias Hollerer, and Sibel Adalii. 2012. Credibility in context: An analysis of feature distributions in twitter. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 293–301. IEEE.

Fernando Perez-Tellez, David Pinto, John Cardiff, and Paolo Rosso. 2011. On the difficulty of clustering microblog texts for online reputation management. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 146–152. Association for Computational Linguistics.

Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning*, 3(3):237–271.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Damiano Spina, Julio Gonzalo, and Enrique Amigó. 2013. Discovering filter keywords for company name disambiguation in twitter. *Expert Systems with Applications*.

Hans van Halteren. 2008. Source language markers in europarl translations. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 937–944. Association for Computational Linguistics.

Carl Vogel, Ger Lynch, Erwan Moreau, Liliana Mamani Sanchez, and Phil Ritchie. 2013. Found in translation: Computational discovery of translation effects. *Translation Spaces*, 2(1):81–104.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In *ICWSM*.

Surender Reddy Yerva, Zoltán Miklós, and Karl Aberer. 2011. What have fruits to do with technology?: the case of orange, blackberry and apple. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, page 48. ACM.