Extracting descriptions of problems with product and services from twitter data

Narendra K. Gupta
AT&T Labs - Research, Inc
Florham Park, NJ 07932 - USA
ngupta @research.att.com

ABSTRACT

There is enough evidence that social media contains timely information that businesses could use to their benefits. In this paper we discuss automatic extraction of descriptions of problems from twitter data. More specifically we present a system that filters tweets related to an enterprise and extracts descriptions of problems with their product/service. First step of this extraction process is to identify the tweets containing such descriptions. We view this as text classification problem. We propose that sentences describing problems can be characterized by their lexical and syntactic structure. Our experiments show that use of such structural features in classification models, results in the F-measure of 0.742. It is a significant improvement over a baseline F-measure of 0.66, obtained by using only word ngram features. Since twitter data is dynamic, classification models have to be adopted to changing nature of problems and language distributions. We describe a simple adaptation scheme, and experimentally demonstrate its effectiveness. Finally we discuss our method to pinpoint the phrases describing the problems in the identified tweets. We show that by using simple syntactic pattern an extraction F-measure of 0.434 is achieved. Considering the noise in the tweeter data this level of performance is quite encouraging.

General Terms

Text classification, Information extraction, Natural Language processing, Machine Learning

1. INTRODUCTION

In the past decade social media has become quite popular and a rich source of variety of information. Twitter [16], is one of the popular on-line social network. It is used by millions of people around the world to remain socially connected to their friends, family members and co-workers by posting short textual messages, called tweet through computers or increasingly through mobile devices. As of September 2010 Twitter had over 145 million registered users and over of 90 million tweets per day [20] and growing fast. One of the important characteristics of twitter data is that it is real time. Whenever some event takes place, people make many tweets about it in real-time. There are a number of studies exploiting this real-time nature of tweets. For example Sakaki et al., [22], describe faster generation of earthquake reports by using tweets as sensors than the traditional broadcast system. Qiu et. al., [19] have concluded that twitter can be a complimentary source of information for monitoring the AT&T network performance. They found that a) tweets about the AT&T network problems were available much before than the incident reports b) for some short term problems tweets were the only source of information.

On social media sites like Twitter, people talk about their personal experiences and preference. It is a rich source of information and can also be used by businesses to gather information about their product and services. Figure 1 shows different types of information that may be useful to a business for different product and services and for different business logistics (marketing, sales, customer service etc). Using social media to mine popular opinions

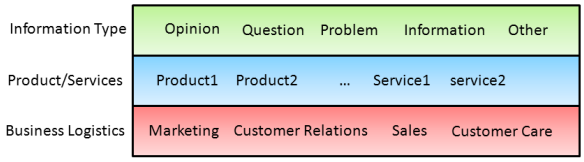

Figure 1: Different types of information useful to a business for different product and services and for different business logistics.

about product and services, political candidates, and even plans and policies from is quite common. As shown in Figure 1 businesses could also learn about the problems consumers face with their products/services, the questions they may have, or even specific information they may have discovered. Examples of tweets containing different types of information for the telecommunication products and services are shown in Figure 2.

In this paper we discuss a method that by using twitter data identifies the problems customers may be facing with the products/services of a company. Our method has three main steps. In the first step tweets containing description of some problem are identified. In the rest of the paper we will refer to such tweets as *problem tweets*. We view problem tweet identification as a text classification problem. We use supervised method to train the classification model and experimentally demonstrate the boost in the identification performance by using the carefully selected feature set. In the second step, phrases describing the problem are extracted from the problem tweets. In the rest of the paper we will refer them as *problem*

Opinion Tweets:

- `Thank God tmobiles roaming partner is at&t.`
- `I've have this phone for 2 months now, and my $100 rebate finally came today. Thanks AT&T, you're awesome.`

Problem Tweets:

- `Anyone else in boston that can't place or receive calls on the network right now?`
- `Computer is not running any better since I talked to them yesturday.`

Question Tweets:

- `Hey AT&T. When are you going to let me tether?`
- `Which is a better cell phone service AT&T or SPRINT?`
- `Any details on that email leak?`

Information Tweets:

- `AT&T doesn't serve our condo community.`
- `Verizon and AT&T have cut the cost of unlimited calling plans by almost a third. Other plans are being discounted as well.`

Figure 2: Examples of tweets containing different types of information for products and services provided by a telecommunications services company.

phrases. We will show that using simple syntactic pattern matching for this purpose yields practically usable results. In the third step we intend to summaries the identified problem phrases by semantic clustering. We have not implemented this step as yet, and will discuss it as a part of our future work. We will use AT&T as an example, and will identify and extract problem phrases describing problems consumers face with AT&T product and services.

Rest of the paper is organized as follows. In section 2 we discuss the previous automation attempted on social media. In section 3 we discuss our approach, which besides pre-processing the tweets, consists of 3 main steps, i.e., problem tweet identification, problem phrase identification and summarization. In section 4 we describe our experiment and their results. In section 5 we describe future work. We conclude the paper in section 6 by summarizing the contributions of this work.

2. RELATION TO PRIOR WORK

Most of the posts on social media, particularly on twitter, are pointless ramblings. Studies [2] have shown that only about 3.6% of these posts contain topics of general interest. Therefore, the first task is to identify the potentially relavant postings for the analysis. These postings are then analyzed to yield information of interest, i.e, opinions about a product, person, or a company or information about an event. Identification of potentially relevant posts on twitter is typically done through keyword spotting. A number of tools, i.e., `socialmention.com`, `sency.com`, `yacktrack.com` and `addictomatic.com` are available to do this filtering. Analysis of the potentially relevant postings is a more difficult task and is to topic of active research.

Mining social media for opinions about products and services [8, 17, 15], and about political candidates or even about some policy/point of view is also very useful. Therefore, a large body of work [5, 12, 18, 14] exists on opinion mining. Since users having problems typically also have negative opinions/emotions, we use some of the same features used in opinion mining to identify the problem tweets, i.e., we use dictionaries of positive and negative opinion words/phrases, and other emotion expressions like emoticons and repeated punctuations in the tweets as features for identification of problem tweets.

There are also many efforts in exploiting real time nature of the twitter in generating reports about natural disasters. In [22]twitter data is used to generate earthquake notifications. To generate such notifications first a set of potentially relevant postings containing hand selected keywords are identified. Presence of such keywords like `earthquake` or `shaking` in a tweet alone does not guarantee the relevance of tweet to earthquake monitoring. For example the tweets `I am in Japan a land of earthquakes` and `They had a minor earthquake 2 weeks back` are not a relevant tweets, because neither of these refer to an earthquake event happening in real time. Similar to our approach, they also identify relevant tweets through text classification. To train the models they use context of occurrences of the hand selected keywords within the tweets as features. This disambiguate the semantics of the such keywords. They also use words referring to time as features. This makes sure the real time nature of the tweet. Using relevant tweets and their meta-data (time and place of origination) they produce a probabilistic spatiotemporal model that can find the center and the trajectory of the event location. In our context time and place of the problem are also important, however, in this work we chose to ignore them.

Another application exploiting real time nature of twitter is described in [25]. In this application twitter data is manually analyzed to identify types of information that must be extracted to improve real time situational awareness in different types of emergency situations. They then manually label the tweets with such information [4] to train information extraction models. In our work we are interested in extracting problem phrases. We also manually analyzed the tweets to extract syntactic patterns of problem phrases occurrences.

A significant body of work also exist in the area of information extraction. This work has been largely spawned from message understanding conferences [1]. For these conferences information extraction is defined as the tasks of identifying the text fragment that answer standard question defined for a document collection. This is similar to our task of identifying the problem phrases within the problem tweets. Extraction rules manually encoded [11] or automatically induced from the tagged data [7, 3, 21, 23] are the backbone for many information extraction system. While some systems are designed to extract information from structured text [13] others are designed for free running text. Such systems try to learn *sequence patterns* that occur inside the phrases of interest and on the left and right boundaries. Words, their parts of speech, and possibly their syntactic category could make up the patterns. Like most learning systems they rely on some regularity of language, still their performance for extracting information from well written free text is very low. For example, for extracting information about management succession Soderland et. al. [23] show both precision and recall in the range of 40-50%. In our context we decided not to follow these approaches mainly because of the noise in the language used in twitter messages, it would be hard to find such sequence patterns.

3. OUR APPROACH

Figure 3 shows the block diagram of our overall approach. In this section we will discuss the pre-processing, problem tweet identification and problem phrase extraction steps.

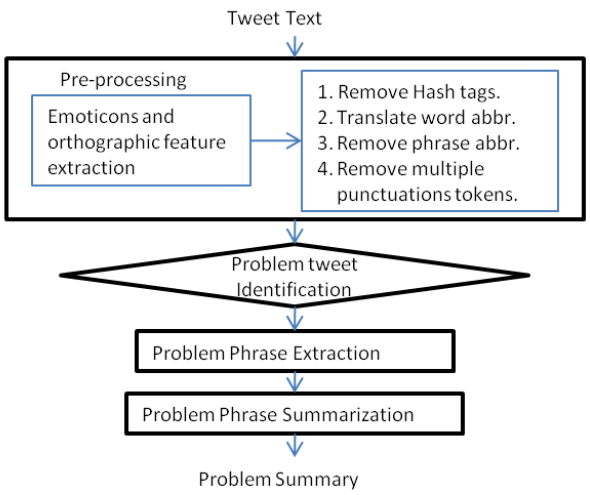

Figure 3: The overall architecture of our solution.

3.1 Pre-processing

For our experiments we selected tweets containing the word "AT&T" in the body. This gives us majority of tweets relating to AT&T products and services. We also filtered out tweets containing any references to a URL. These are mostly tweets from companies pushing their advertisements.

Twitter data is very noisy. Because of message size limitations it is full of abbreviations and other orthographic artifacts. Tweets rarely contain well formed language. This poses many challenges for natural language processing. We describe some of those challenges and also how we handle them during pre-processing.

3.1.1 Hash tags

Majority of messages contain hash tags. They are often formed by concatenating words that explain their meaning for example, `#retweetthisif` and `#truefacts`. Hash tags are sometime used as part of the sentence and other time only as a referential labels. In `at&t announced when its 1st #droid will be available` the hash tag is part of the sentence and in `You know him as Dave from AT&T. #bigbangtheory` it is used as a label.

We remove most of the hash tags from our data and retain only those that refer to the company (in our case AT&T) or have words that are indicative problem tweets, for example `fail, break, fix, repair` or `outage` embedded in them.

3.1.2 Word Abbreviation

Often words are arbitrarily abbreviated. Humans are able to parse them, but it would be difficult to automatically deduce the correct words from such abbreviations. An example of such a message is `u aint lyin. at&t plans be gettin over tho.`

To deal with such abbreviations we selected high frequency (greater than 3) non-dictionary words from 10000 randomly selected tweets. We manually went through these words and created a dictionary of their translations with 65 entries. Example of entries in this dictionary is shown in Table 1. During the pre-processing word abbreviations in the dictionary are replaced by their translations.

Abbreviation	Translations	Abbreviation	Translations
ystdy	yesterday	rly	really
sux	sucks	shld	should
cnt	can't	phn	phone
ur	your	ain	Is not
wrkn	working	fukd	f**ked

Table 1: Example abbreviations and their translations.

3.1.3 Phase Abbreviations

At least 3 different kinds of phrase abbreviations are encountered.

1. Filler phrases like `btw` for `by the way`, `imo` for `in my opinion`.
2. Emotion expressions like `omg` for `oh my God` and `lol` for `laughing out loud`m
3. Meaning phrases which contribute to the meaning of tweet like `imma` for `I am going to` and `wanna` for `want to`.

Some phrase abbreviations are universal and can be decoded while others are made-up on the fly and are difficult to decode.

In 10000 randomly sampled tweets we identified a list of 23 high frequency (greater than 3) filler and emotional phrase abbreviations. During the pre-processing occurrences of these phrase abbreviations are deleted. Phrase abbreviations of meaningful phrases are handled through the word abbreviation dictionary described above.

3.1.4 Made up word

Made up words are mostly used to emphasize something or to express emotions, i.e., in `Dish network has a gr8 promo, and so do AT&T ... yippee!!!` and `i kno! takin shots at T.I. too... owwee`. While words like `yippee` and `owwee` in these tweets express emotions, other examples like `splendiferous` and `massassssive` are purposely deformed to extenuate their meaning. How to deal with such made-up-words in itself is research topic. In this work, we did not focus on them and left them as is in the text.

3.1.5 Repeated punctuations

Repeated punctuations are mostly used to express emotions. Some examples of such expressions are: `Service in Pikeville is pathetic right now!!!!`,`What's the deal with this sh*tty service?????`. After their presence is recorded, such repeated punctuations are replaced by a single punctuation mark.

3.1.6 Emoticons

Twitter data is fraught with emoticons. People do not always use standard emoticons. For example, to express anger, emoticons like `>:/, :-/, :(«, >:[` etc are used. Such creativity is bound to cause recall and precision problems in extracting the emoticons.

Even though there are emoticons for many different emotions, during pre-processing we only extract emoticons expressing happiness, sadness and anger. After their presence is recorded, all instances emoticons are removed from the text for further processing.

3.2 Problem Tweet Identification

We view the problem tweet identification as a text classification problem. We trained Maximum Entropy classifier using word ngrams and other features described below.

3.2.1 Sentiment Features

Users having some kind of problem, often express negative sentiments, either by negative emotions like being angry or upset or by simply stating their negative opinion. To capture these we extract following features from the tweets.

1. **Emoticons:** Three binary features are used to indicate presence or absence of happy, sad and angry emoticons in the tweets.

2. **Orthographic:** Four binary features are used to indicate presence/absence of token consisting of repeated exclamations (!), interrogation (?), periods (.) and dollar ($) signs in the tweets.

3. **Positive and negative Sentiment:** Two binary features are used to indicate presence/absence of phrases expressing positive/negative sentiment in the tweets. To extract these features we use dictionaries of positive/negative phrases. This dictionary contained 811 phrases expressing positive sentiment and 529 phrases expressing negative sentiment. It was populated from samples of customer emails in the context of customer service [10], and samples of customer reviews [9] of products and services. Examples of these phrases are shown in Figure 4.

Positive Phrases:
`incredible, entertaining. pleasing, inexpensive, perfection, snappy, wow, real deal, n't bad, damn good, worth it, top notch, rave about, never fail, better than average, have a blast, check it out, a must try, to die for, best keep secret, hand down the best, will not be sorry, feel right at home, right on the money, out of this world, what else can you ask for, as good as can be, n't get any better than this`
Negative Phrases:
`abysmal, malicious, inferior, obnoxious, unwelcome, fearful, below average, below par, mess up, nothing spectacular, mediocre at best, not as good, left something to be desire, run of the mill, not a big fan, nothing to write home about, not live up to the hype`

Figure 4: Example of positive and negative phrases.

3.2.2 Syntactic features

Users describe the product/service problems in many ways. Some of these can be characterized by specific syntactic pattern. Through manual analysis of a few hundred tweets we have identified following syntactic patterns characterizing the problem tweets. For each of these patterns we use a binary feature indicating their presence/absence in the tweet.

Verbs: A common way to describing the problem is by explaining what is happening. Such a description will have a verb phrase with specific head verbs. For example in `My network crashes frequently` the head of the verb phrase (VP) is `crashes` and in `preorder system is broken` it is `broken`. Clearly some verbs are of general nature and can be used across the domain to describe a problem, while others are domain specific for example the verb `freeze` in `My computer freezes often` is specific to the computer domain. We have identified 9 *happening verbs* that with high probability are used across the domain to describe a problem. These verbs are `fail, crash, overload, trip, fix, mess, break, overcharge` and `disrupt`.

Another very common way of describing the problem is by explaining what is not happening; for example in `my internet is not working` or `my internet has stopped working` etc. We identified 10 *not happening verbs*. These are `work, function, connect, get, perform, receive, send, run, stable` and `respond`. These verbs are often used, along with negation, or with head verb `stop`, `refuse` or `cease` to describe what is not happening.

To extract this feature, we test if the tweet contains a verb phrase with a head verb from the list 9 happening verbs, or if it contains description of what is not happening with a head verb from the list of 10 not happening verbs. In the later case either the presence of negation of the verb or open clausal complement dependence of the verb on `stop, refuse or cease` is also tested. In our implementation presence of negation is tested through the dependency `neg(verb, not)`[1] and open clausal complement dependence is tested through `xcomp("stop|refuse|cease", verb)`.

Softer Verbs: These are like the happening verbs mentioned above but are very often also used in context other than problem description. To extract this feature, we test if the tweet contains a verb phrase with head verb being one of `die, drop, bite, f**k, trouble, foil`.

Noun: Another way to describe a problem is by using noun phrases with specific head nouns. For example in `we have an internet failure` head of the noun phrase is `failure` and in `we are having a 3G outage` it is `outage`. To extract this feature, we test if the tweet has a noun phrase headed with one of the *problem nouns*. We used `crash, failure, issue, outage, problem` and `trouble` as problem nouns.

Phrases: A number of phrases are frequently used to describe the problems, i.e., `screwed up`, `hang up`, `knock off`, `knocked out`, `acting up` etc. In these phrases propositions act as particle of the verb. Since particles do not always immediately follow the verbs, for example in `They have screwed my service up`, phrase spotting will miss many occurrences of such phrases. To

[1] In this paper we will use the dependencies generated by Stanford parser. See [6] for details.

spot them we check for `prt(verb, proposition)` dependency between verbs and propositions listed in table 2.

Verbs	f**k, hang, screw, cut, knock, act
Propositions	up, off, out

Table 2: Verbs and propositions in particle dependencies.

Besides phrases with particles, other phrases are also used to describe a problem. In our analysis we found multiple occurrences of phrases with `act` and `behave` but without any particles. For examples `acting funky` in tweet `My internet is acting funky today`. To test occurrences of such phrases in a tweet we test if the parse tree of the tweet had a verb phrase with `act` or `behave` as its head and with either an adjective (ADJP) or adverb (ADVP) phrase as its sub-constituent. This is illustrated by parse trees in Figure 5. The above mentioned pattern is found in parse 1 which contains the problem description and not in parse 2 which does not contain any problem description.

Parse1: `My internet is acting really weird`

```
(S acting
 (NP internet (PRP$ My)(NN internet))
 (VP acting (VBZ is)
   (VP acting (VBG acting)
     (ADJP weird (RB really)(JJ weird)))))
```

Parse2: `He act like my boss`

```
(S act-2
  (NP He-1(PRP He))
  (VP act-2(VBP act)
    (PP (IN like)
      (NP boss-5(PRP$ my)(NN boss)))))
```

Figure 5: Parse trees of two sentences with and without problem descriptions. Head words of constituents are inserted between constituent name and its sub-structure.

3.3 Problem Phrase Extraction

To extract problem phrases from problem tweets we use the syntactic patterns described in previous sub-section as triggers. Only when the presence of those syntactic patterns is detected, syntactically related phrases are assumed to be the problem phrases as described below.

3.3.1 Verb or Softer Verb patterns

If a problem tweet contains Verb or Soft Verb patterns then we assume that either the subject or the object of those verbs to be the problem phrase. In this work we select the subject (noun phrase) of the verb unless it is composed of single pronoun, in which case we extract the direct object (noun phrase) as problem phrase. Extracting subject or the object of a verb in complex sentences requires attention to clausal complements and active/passive form of the sentence. In our implementation we do consider these as well. In Table 3 we provide examples of problem tweets and have highlighted the triggering verb pattern and the corresponding problem phrase.

3.3.2 Noun Patterns

My phone *can't connect*
My Internet has been *messed up*
my Ball on my bb *don't work*
Network *failed*
my unlmtd text plan was accidentally *dropped* last month
They can *fix* **my reward Problem .**
Dropped **my phone** into a lake while riding my jetski

Table 3: Example of problem phrases (bold) extracted as subject/object of triggering verb patterns (italics).

To extract problem phrases from tweets containing noun patterns we select the highest noun phrase (the NP node) in the parse tree with one of the problem nouns listed previous subsection, as its lexical head. Examples of such problem tweets highlighted with the problem phrases and their triggering noun pattern are shown in Table 4.

They are having **bandwidth** ***issues***
network ***outage*** is affecting all of Seattle
I have never had **so much** ***trouble*** **trying to give money**

Table 4: Example of problem phrases (bold) extracted by using noun patterns. Triggering noun patterns in these examples are sub-strings of problem phrases.

3.3.3 Phrase Patterns

We deal with phrases shown in Table 2 in the same way as the verbs and the soft verbs with an additional check to make sure that the propositions listed in Table 2 have particle dependency with the verbs. Phrases with `act` and `behave` that do not have particle dependency are also dealt with in the same way after existence of their syntactic patterns is confirmed. In our analysis, we found many phrases where the word `down` was used to describe a problem, i.e., `shut down`, `went down` and `are down`. These phrases are not specific to problem description and are also used as frequently in other context i.e., `I went down to check on him`. However, in a problem tweet we assume that such phrases are used only to describe the problems. To isolate the problem phrase in such cases we first search the parse tree for a constituent of type ADJP, ADVP or PRT (adjective adverb or particle phrase respectively) with lexical head "down". If the parent of this constituent is a verb phrase (VP) we extract subject or the object of its lexical head verb. On the other hand if the parent of this constituent is a sentence (S) we extract the first noun phrase (NP) from its constituent list.

Top portion of Figure 6 shows an example when the parent of ADVP (with lexical head `down`) is a VP with lexical head `are`. Dependency parse shows the subject of verb `are` is `Lines`. Bottom part of this figure shows another example when the parent of ADJP (with lexical head `down`) is a S with noun phrase (NP) `some circuits`.

4. EXPERIMENTS

For experimental purposes we randomly selected about 10000 tweets from Jan 2010 to August 2010, containing a reference to AT&T. We manually labeled this data with two class labels "Problem"

Tweet: `Lines are down since morning.`

```
Syntactic Parse:
(S are
  (NP Lines (NNP Lines))
  (VP are (VBP are)
    (ADVP down (RB down))
    (PP (IN since)
      (NP morning (NN morning)))))
Dependency Parse:
nsubj(are-2, Lines-1)
advmod(are-2, down-3)
prep_since(are-2, morning-5)
```

Tweet: `They have some circuits down.`

```
Syntactic Parse:
(S have-2
  (NP They (PRP They))
  (VP have (VBP have)
    (S down
      (NP circuits (DT some)(NNS circuits))
      (ADJP down-5(RB down)))))
```

Figure 6: Examples of phrase pattern with word `down`**.**

and "No-problem". To assess the labeling consistency we also measured inter-annotator agreement by letting two labelers label 2000 randomly selected tweets. A Kappa value of 0.806 shows that the labeling task is relatively hard. Most of the disagreement arises form the ambiguity between expressions of negative opinions and descriptions of problems. For example in `My upload speed is way too slow !!` negative opinion about the upload speed is being expressed and in `Since yesterday my upload speed has become very slow !!` a problem with upload speed is being reported. Other disagreements are due to the different interpretations of what a problem tweet is. For example writer of the tweet `I just got the service interruption notice from AT&T. It pisses me off` has a problem, but it is not a problem with the service that the company can fix.

To perform our experiments we randomly partitioned the data in training and test set of 8912 and 980 tweets respectively, containing 8.89% and 8.88% problem tweets respectively. Below we show experimental results on problem tweet identification and problem phrase extraction using this data.

4.1 Problem Tweet Identification

In our first experiment we trained a maximum entropy classification model using only unigram words as features, and tested it using the test data. The performance of this baseline model is annotated by "Uni-Gram Words" in Figure 7. Recall and precision are calculated at various confidence thresholds values. As can be seen the best identification performance of 0.66 F-value is obtained.

In our second experiment in addition to unigram word features, we also used all the sentiment and syntactic features described in section 3. The results of this experiment is annotated by "Additional Features" in Figure 7. In this case best F-value of 0.742 is obtained, which a significant improvement over baseline. Error analysis revealed that classifier has difficulty identifying problem tweets

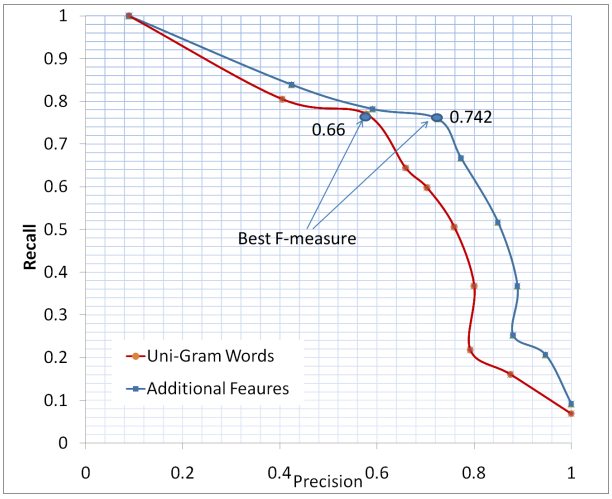

Figure 7: Problem tweet identification performance.
sample

where problems are described in a manner that requires deep domain knowledge. An example of such a tweet is `i juss got a text messages that were sent more than 30 mins ago.. wth?`. In further discussion we will refer to such tweets as *domain problem tweets* To identify them using statistical classifiers, an order of magnitude more training data must be labeled. In mining social media data where nature of problems changes with time such labeling is not feasible. In spite of this, we think that with our model many of the problems described in domain problem tweets will also be eventually extracted. For example, if the problem with text messages described in the above tweet is faced by many users, it is very likely that some user will generate the tweet along the lines of `my text messaging is not working. I just got messages that were sent an hour back`, which does not require the deep domain knowledge for identification as problem tweet.

4.2 Adaptation of Classification Models

Twitter data is very dynamic. Both the types of problems consumer face and the language they use to express them change over time. To demonstrate this we tested the classification model described in section 4.1 (trained on randomly drawn data from Jan. to Aug. 2010) on two additions sets of test data. One randomly drawn from Jan. to March 2011 and the other drawn from April 26, 2011. Table 5 clearly shows the classifier performance degrades with time.

Data ID	Period Drawn	Num. Examples	Prec.	Rec.	F
T1	Jan.-Aug.2010	980	0.725	0.759	0.742
T2	Jan.-Mar.2011	1000	0.864	0.559	0.679
T3	Apr. 26 2011	1000	0.7	0.677	0.689

Table 5: Performance of past model on test data from later period.

To overcome such temporal degradation of classification performance, models must be periodically retrained on the current training data. Instead of completely relabeling the freshly drawn training data every time, we use active learning technique [24] and selected only a fraction of training examples to be examined. To do this we follow the steps shown in Figure 8. We classify the freshly

drawn tweets using the current classification model. For each example t_i our model (maximum entropy) outputs a confidence value between 0-1, which can be interpreted as the probability of the tweet being a problem tweet, i.e. $p(prob|t_i)$. We then select all the examples with $p(prob|t_i) > Th$ and label them. These labeled examples together with previously used training data are used to train the new model.

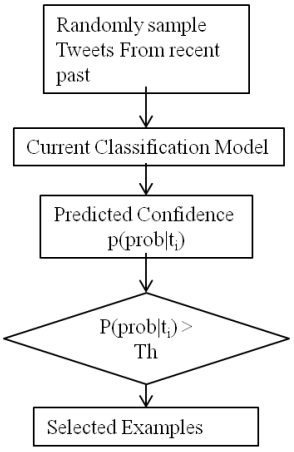

Figure 8: Selection of examples to be labeled for model adaptation.

In our experiments at each adaptation iteration we drew 10000 fresh tweets and used $Th = 0.1$ to select the examples for labeling. Our choice of selection threshold may seem very small but it only results in selection of approximately 10% of tweets for labeling. Since best classifier performance (best F-measures shown in Figure 7) is obtained at $p(prob|t_i) > 0.3$, our selection threshold forces us to examine and label all the predicted problem tweets. This helps in inceasing the precision of the new model. It also forces us to examine small fraction of predicted non-problem tweets that are closer to the decision boundary. This helps improve the recall.

Table 6 shows the results of two adaptation iterations. To adapt the model for iteration 1 only 712 fresh examples from Jan-March 2011 were selected for labeling. A new model was trained using original 8910 training examples and 712 freshly labeled example. As can be seen from the table the performance of this model on test data T2 which is from Jan.-March 2011 is improved from F-measure of 0.679 to 0.724. Even the performance on the test data T3 which is from a later period is improved from 0.689 to 0.709. For iteration 2 additional 1009 examples were selected from April 1 to April 25, 2011, and the new model improved the performance on test data T3 which is from April 26 from F-measure of 0.709 to 0.775.

Adaptation iteration	Training data count	Test data ID	Prec.	Rec.	F-measure
1	9622	T2	0.875	0.618	0.724
		T3	0.7	0.718	0.709
2	10631	T3	0.756	0.795	0.775

Table 6: Modal adaptation iterations.

4.3 Problem Phrase identification

To evaluate the problem phrase identification procedure described in previous section we labeled the problem phrases in the problem tweets of the test data, and compared the results with those extracted by our method. We obtained a precision of 0.434 and recall of 0.206 (F-Value of 0.279). Error examination revealed that to compute precision and recall, exact matching between extracted and reference phrases is too restrictive. In Table 7 we show some randomly selected examples of the extracted and reference phrases which overlap but do not match exactly. As can be seen, in almost all the examples the mis-match is due to the presence or absence of one or two words, still the extracted phrases convey almost the same information as the reference phrases. To make a more realistic comparison therefore, we relaxed matching criterion of phrases and require an overlap of only 50% of the words to consider a match. Using this relaxed matching, a precision of 0.663 and recall of 0.323 (F-Value of 0.434) was obtained. This is the upper bound of over-all performance as it assumes correct identification of problem tweets. This low performance of our method is due to the fact that our method relies on correct syntactic parse of the noisy tweets. Furthermore, none of the problem phrases from the *domain problem tweets* were extraced. Ignoring such tweets in evaluation results in a precision of 0.663 and recall of 0.542 (F-Value of 0.596). While this level of performance is useful, still

Reference Phrases	Extracted Phrases
my network	Dang my network
network outage	a network outage
iphone coverage	just no iphone coverage
freaking signal	the freaking signal
New DSL modem	New DSL
damn Internet	this damn Internet
my phone	great now my phone
service	extreme service
skype	skype tonight ATT

Table 7: Random examples of overlapping but not exactly matching extracted and reference phrases.

better algorithms are needed to extract problem phrases. In particular, to decide whether to extract the subject or the object of the verb and also to reject those extractions caused by bad syntactic parses or noisy tweets. In next section we describe some ideas we intend to pursue as part of the future work.

5. FUTURE WORK

In this paper we have presented some lexical and syntactic patterns that characterize the descriptions of problems one faces with products/services offered by a company. These patterns were identified by analysis of tweets referring to AT&T. In future we would like to test how useful these patterns are for problem tweet identification in other domains like computers, televisions etc. We would like to analyze data from those domains and see if there are other general syntactic patterns used to describe problems and investigate how their use affect the performance on AT&T data.

Problem phrase identification described in this paper is completely based on syntactic patterns and has poor performance. Large number of errors take place because of bad parses and extraction of wrong subject/object of the head verbs, particularly from complex sentences. In future work we will use a ranking model where all candidate problem phrases are ranked with respect to their relevance. We will use supervised methods to train the model and will

also benefit from prior knowledge of products and services of interest.

Large amount of noise in twitter data handicaps application of any NLP technique. In our experience multiple usage of hash tags, arbitrary phrase abbreviations and made up words are the most difficult types of noise to deal with. To alleviate these problems we are also exploring twitter text normalization through the language models.

In this paper we have not discussed the final step of our approach, i.e., problem phrase summarization as shown in Figure 3. It requires semantic clustering of the problem phrases. Besides using the word-net synsets as features to compute similarity, we also intend to use other features like tweet origination location and time, words representing products/service in the domain of interest and re-tweet counts as features.

6. CONCLUSION

In this paper we proposed a method for mining the twitter data for the problems customers may be facing with products/services of a company. We divided the task in 3 steps. We discussed the first two steps viz. problem tweet identification and the problem phrases extraction. We treated the problem tweet identification as classification task and devised new features based on lexical and syntactic patterns often used to express the problems with products and services. We demonstrated that by using these features, identification performance is boosted from a baseline f-measures of 0.66 to 0.742. We also showed that by using active learning technique classification models can be adapted to dynamic nature of twitter data by minimal amount of labeling. To extract problem phrases, we used the same syntactic patterns and obtained an F-measure of 0.434. With the lexical, grammatical and orthographical noise present in the tweets this level of performance is quite encouraging. In future work we intend to improve it further .

7. REFERENCES

[1] Proceedings of the sixth message understanding conference, 1995.

[2] P. Analytics. Pear analytics: Twitter study-august-2009. http://www.scribd.com/doc/18548460/ Pear-Analytics-Twitter-Study-August-2009, 2009.

[3] M. E. Califf and R. J. Mooney. Bottom-up relational learning of pattern matching rules for information extraction. *Journal of Machine Learning Research*, pages 177–210, 2003.

[4] W. J. Corvey, S. Vieweg, T. Rood, and M. Palmer. Twitter in mass emergency: what nlp techniques can contribute. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 23–24, Stroudsburg, PA, USA, 2010.

[5] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*, pages 519–528, 2003.

[6] M.-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 2006 LREC*, 2006.

[7] D. Freitag. Toward general-purpose learning for information extraction. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, pages 404–408. Association for Computational Linguistics, 1998.

[8] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *In KDD*, pages 419–428, 2005.

[9] N. Gupta, G. Di Fabbrizio, and P. Haffner. Capturing the stars: Predicting ratings for service and product reviews. In *Procceedings of the Workshop on Semantic Search at NAACL-2010*, Los Angeles, CA, USA, 2010.

[10] N. Gupta, M. Gilbert, and G. Di Fabbrizio. Emotion detection in email customer care. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text at NAACL-2010*, Los Angeles, CA, USA, 2010.

[11] J. R. Hobbs, D. E. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In E. Roche and Y. Schabes, editors, *Finite State Devices for Natural Language Processing*. MIT Press, Cambridge, MA, 1996.

[12] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the COLING 2004*, 2004.

[13] N. Kushmerick, D. S. Weld, and R. Doorenbos. Wrapper induction for information extraction. In *Proceedings of the fifth internationsl joint conference on Artificial Intelligence*, pages 729–737, 1997.

[14] J. Liu and S. Seneff. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the 2009 EMNLP*, pages 161–169, Singapore, 2009.

[15] C. Macdonald and I. Ounis. The trec blog06 collection : Creating and analysing a blog test collection. In *DCS Technical Report TR-2006-224*, University of Glasgow, 2006.

[16] S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica, and R. Magoulas. Twitter and the micro-messaging revolution: Communication, connections, and immediacyŬ140 characters at a time. http://www.weigend.com/files/teaching/ischool/2008/ readings/OReillyTwitterReport200811.pdf, 2008.

[17] K. Nigam and M. Hurst. Towards a robust metric of opinion. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Standford, CA., 2004.

[18] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 EMNLP*, pages 79–86, Philadelphia, PA, USA, 2002.

[19] T. Qiu, J. Feng, Z. Ge, J. Wang, J. Xu, and J. Yates. Listen to me if you can: Tracking user experience of mobile network on social media. In *Proceedings of IMC'1*, Melbourne, Australia, 2010.

[20] L. Rao. Twitter seeing 90 million tweets per day, 25 percent contain links. http://techcrunch.com/2010/09/14/twitter-seeing-90-million-tweets-per-day/, 2010.

[21] E. Riloff. An empirical study of automated dictionary construction for information extraction in three domains. *ARTIFICIAL INTELLIGENCE*, 85:101–134, 1996.

[22] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the WWW2010*, Raleigh, North Carolina, USA, 2010.

[23] S. Soderland, C. Cardie, and R. Mooney. Learning information extraction rules for semi-structured and free text. In *Machine Learning*, pages 233–272, 1999.

[24] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *JOURNAL OF MACHINE LEARNING RESEARCH*, pages 999–1006, 2000.

[25] S. Vieweg, A. L. Hughes, K. Starbir, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 2010 CHI*, pages 1079–1088, Atlanta, Georgia, USA, 2010.

Acknowledgments

Author wishes to thank Mazin Gilbert for providing continuing support in carrying out this work.