

## An Iterative Approach for Identifying Complaint Based Tweets in Social Media Platforms (Student Abstract)

Gyanesh Anand,<sup>1\*</sup> Akash Gautam,<sup>1\*</sup> Puneet Mathur,<sup>2</sup> Debanjan Mahata,<sup>3</sup> Rajiv Ratn Shah,<sup>1</sup> Ramit Sawhney<sup>4</sup>

<sup>1</sup>MIDAS, IIIT-Delhi {gyanesh16039, akash15011, rajivrtn}@iiitd.ac.in,

<sup>2</sup>University of Maryland, College Park puneetm@cs.umd.edu,

<sup>3</sup>Bloomberg, New York, U.S.A. dmahata@bloomberg.net,

<sup>4</sup>Netaji Subhas Institute of Technology ramits.co@nsit.net.in

### Abstract

Twitter is a social media platform where users express opinions over a variety of issues. Posts offering grievances or complaints can be utilized by private/ public organizations to improve their service and promptly gauge a low-cost assessment. In this paper, we propose an iterative methodology which aims to identify complaint based posts pertaining to the transport domain. We perform comprehensive evaluations along with releasing a novel dataset for the research purposes<sup>1</sup>.

### Introduction

With the advent of social media platforms, increasing user base address their grievances over these platforms, in the form of complaints. According to (Olshtain and Weinbach 1985), *complaint is considered to be a basic speech act used to express negative mismatch between the expectation and reality*. Transportation and its related logistics industries are the backbones of every economy<sup>2</sup>. Many transport organizations rely on complaints gathered via these platforms to improve their services, hence understanding these are important for: (1) linguists to identify human expressions of criticism and (2) organizations to improve their query response time and address concerns effectively.

Presence of inevitable noise, sparse content along with rephrased and structurally morphed instances of posts, make the task at hand difficult (Shah and Zimmermann 2017). Previous works (Meinl 2013) in the domain of complaint extraction have focused on static datasets only. These are not robust to changes in the trends reflected, information flow and linguistic variations. **We propose an iterative, semi-supervised approach for identification of complaint based tweets, having the ability to be replicated for “stream of information flow”**. The preference of a semi-supervised approach over supervised ones is due to the stated reasons: (a) the task of isolating the training set, make

supervised tasks less attractive and impractical and (b) imbalance between the subjective and objective classes lead to poor performance.

### Proposed Methodology

We aimed to mimic the presence of sparse/noisy content distribution, mandating the need to curate a novel dataset via specific lexicons. We scraped 500 random posts from recognized transport forum<sup>3</sup>. A pool of 50 uni/bi-grams was created based on tf-idf representations, extracted from the posts, which was further pruned by annotators. Querying posts on *Twitter* with extracted lexicons led to a collection of 19,300 tweets. In order to have lexical diversity, we added 2500 randomly sampled tweets to our dataset. In spite of the sparse nature of these posts, the lexical characteristics act as **information cues**.

Figure 1 pictorially represents our methodology. Our approach required an initial set of *informative tweets* for which we employed two human annotators annotating a random sub-sample of the original dataset. From the 1500 samples, 326 were marked as *informative* and 1174 as *non informative* ( $\kappa = 0.81$ ), discriminated on this criteria: ***Is the tweet addressing any complaint or raising grievances about modes of transport or services/ events associated with transportation such as traffic; public or private transport?*** An example tweet marked as *informative*: “No, metro fares will be reduced ???, but proper fare structure needs to be presented right, it’s bad !!!”.

We utilized tf-idf for the identification of initial “seed phrases” from the curated set of *informative tweets*. 50 terms having the highest tf-idf scores were passed through the complete dataset and based on sub-string match, the *transport relevant tweets* were identified. The redundant tweets were filtered based on the cosine similarity score. **Implicit information indicators** were identified based on *domain relevance score*, a metric used to gauge the coverage of n-gram (1,2,3) when evaluated against a randomly created pool of posts.

We collected a pool of 5000 randomly sampled tweets different from the data collection period. The rationale behind

\*Equal Contribution

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup>The dataset can be found at <https://github.com/midas-research/transport-complaint-detection>

<sup>2</sup><https://www.entrepreneur.com/article/326552>

<sup>3</sup><https://www.theverge.com/forums/transportation>

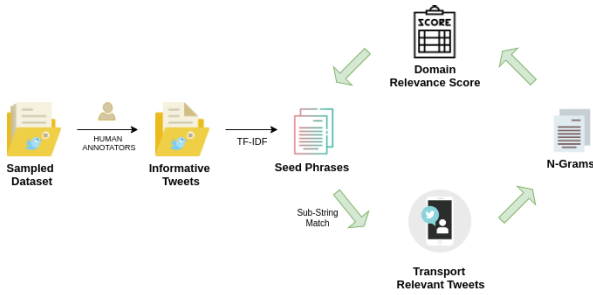


Figure 1: Pictorial representation of the proposed pipeline.

having such a metric was to discard commonly occurring n-grams normalized by random noise and include ones which are of lexical importance. We used terms associated with high domain relevance score (threshold determined experimentally) as “seed phrases” for the next set of iterations. The growing dictionary augments the collection process. The process ran for 4 iterations providing us 7200 *transport relevant tweets* as no new lexicons were identified. In order to identify linguistic signals associated with the *complaint* posts, we randomly sampled a set of 2000 tweets which was used as training set, manually annotated into distinct labels: *complaint relevant* (702) and *complaint non-relevant* (1298) ( $\kappa = 0.79$ ). We employed these features on our dataset.

**Linguistic markers.** To capture linguistic aspects of complaints, we utilized Bag of Words, count of POS tags and Word2vec clusters.

**Sentiment markers.** We used quantified score based on the ratio of tokens mentioned in the following lexicons: MPQA, NRC, VADER and Stanford.

**Information specific markers.** These account for a set of handcrafted features associated with *complaint*, we used the stated markers (a) Text-Meta Data, this includes the count of URL’s, hashtags, user mentions, special symbols and user mentions, used to enhance retweet impact; (b) Request Identification, we employed the model presented in (Danescu-Niculescu-Mizil et al. 2013) to identify if a specific tweet assertion is a request; (c) Intensifiers, we make use of feature set derived from the number of words starting with capital letters and the repetition of special symbols (exclamation, questions marks) within the same post; (d) Politeness Markers, we utilize the politeness score of the tweet extracted from the model presented in (Danescu-Niculescu-Mizil et al. 2013); (e) Pronoun Variation, these have the ability to reveal the personal involvement or intensify involvement. We utilize the frequency of pronoun types {*first, second, third, demonstrative and indefinite*} using pre-defined dictionaries.

From the pool of 7200 transport relevant tweets, we sampled 3500 tweets which were used as the testing set. The results are reported in Table 1 with 10 fold cross-validation. **With increasing the number of iterations, the pool of “seed phrases” gets refined and augments the selection of transport relevant tweets.** The proposed pipeline is tailored to identify complaint relevant tweets in a noisy scenario.

Model	Accuracy(%)	F1-score
<b>Linguistic Markers</b>		
<b>Bag-of-Words</b>	<b>75.3</b>	<b>0.71</b>
POS Tags	70.1	0.66
Word2Vec cluster	72.1	0.67
<b>Sentiment Markers</b>		
Sentiment-MPQA	68.2	0.61
Sentiment-NRC	67.9	0.59
Sentiment-VADER	68.0	0.62
<b>Sentiment-Stanford</b>	<b>68.7</b>	<b>0.63</b>
<b>Information Specific Markers</b>		
Text Meta-Data	69.3	0.62
Request Identification	70.1	0.66
<b>Intensifiers</b>	<b>72.5</b>	<b>0.67</b>
Politeness Markers	70.4	0.63
Pronoun Variations	69.6	0.65

Table 1: Performance of various linguistic, sentiment and information specific features on our dataset. **Classifier utilized Logistic Regression (Elastic Net regularization)**, as it gave the best performance as compared to its counterparts.

## Results

Table 1 reflects that the BOW model provided the best results, both in terms of accuracy and F1-score. The best result achieved by a sentiment model was the Stanford Sentiment (0.63 F1-score), with others within the same range and linguistic-based features collectively giving the best performance.

## Conclusion and Future Work

In this paper, we presented a novel semi-supervised pipeline along with a novel dataset for identification of complaint based posts in the transport domain. **The proposed methodology can be expanded for other fields by altering the lexicons used for the creation of information cues.** There are limitations to this analysis; we do not use neural networks which mandate a large volume of data. In the future, we aim to identify demographic features for identification of complaint based posts on social media platforms.

## References

- Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, 307–318. ACM.
- Meinl, M. E. 2013. *Electronic complaints: an empirical study on British English and German complaints on eBay*, volume 18. Frank & Timme GmbH.
- Olshain, E., and Weinbach, L. 1985. *Complaints-A study of speech act behavior among native and nonnative speakers of Hebrew*. Tel Aviv University.
- Shah, R., and Zimmermann, R. 2017. *Multimodal analysis of user-generated multimedia content*. Springer.