Pixie: Preference in Implicit and Explicit Comparisons

Amanul Haque and Vaibhav Garg and Hui Guo and Munindar P. Singh
Department of Computer Science
North Carolina State University
Raleigh, NC 27695, USA
{ahaque2, vgarg3, hguo5, mpsingh}@ncsu.edu

Abstract

We present Pixie, a manually annotated dataset for preference classification comprising 8,890 sentences drawn from app reviews. Unlike previous studies on preference classification, Pixie contains implicit (omitting an entity being compared) and indirect (lacking comparative linguistic cues) comparisons. We find that transformer-based pretrained models, fine-tuned on Pixie, achieve a weighted average F1 score of 83.34% and outperform the existing state-of-the-art preference classification model (73.99%).

1 Introduction

Online user reviews contain a cornucopia of information on user expectations about a product. Users often express their opinions on a product by comparing it against competitors. Understanding preferences in natural language is crucial in capturing user’s opinions and expectations. Previous studies show that app reviews include rich insights about user expectations and problems of mobile apps that are valuable for app developers (Palomba et al., 2015; Maalej and Nabil, 2015; Guo and Singh, 2020). We found that app reviews often include comparative sentences, from which we can determine a reviewer’s preferences.

Identifying the preferred entity from an app review involves (1) Comparative Sentence Identification (CSI) (Jindal and Liu, 2006), i.e., identifying sentences that contain a comparison, and (2) Comparative Preference Classification (CPC) (Ganapathibhotla and Liu, 2008; Panchenko et al., 2019), i.e., identifying the preferred entity in a comparative sentence. We focus on the second task.

Prior work on CPC focuses on explicit comparisons, where all compared entities are explicitly mentioned. Extracting comparative sentences by matching keywords or patterns (Jindal and Liu, 2006; Li et al., 2017; Feldman et al., 2007) over-

<table>
<thead>
<tr>
<th>Sentence</th>
<th>App</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>Bye Uber, hello Lyft.</td>
</tr>
<tr>
<td>S_2</td>
<td>Does this app really need to be 260 MB when the Marriott app is only 47 MB?</td>
</tr>
<tr>
<td>S_3</td>
<td>Beats the pants off pandora.</td>
</tr>
<tr>
<td>S_4</td>
<td>I think that it’s a lot more fun than temple Run.</td>
</tr>
</tbody>
</table>

Table 1: Example comparative sentences from reviews.

looks indirect comparisons which lack comparative quantifiers and adjectives.

Staab and Hahn (1997) identify omitted complement as a comparative sentence type that has been overlooked by prior research. An omitted complement refers to one of the entities under comparison that is omitted but can be inferred based on the context. We have found that comparative sentences in user-generated text such as reviews sometimes imply instead of explicitly mentioning the target entity being reviewed (e.g., S_3 in Table 1). Comparisons in reviews often lack comparative linguistic cues, such as comparative quantifiers, adjectives, or structures (i.e., indirect, e.g., S_1 in Table 1). Such sentences are comparative by virtue of expressing a preference and are common in reviews but have been understudied by prior research.

We present Pixie (Preference in Implicit and Explicit Comparisons), a dataset for preference classification, created from online user reviews. As shown in Table 1, Pixie includes indirect comparisons (i.e., sentences lacking comparative linguistic cues, e.g., S_1) and implicit comparisons (omitting compliments, i.e., mentioning only one entity being compared, e.g., S_3) in addition to direct comparisons (comparing entities with a direct comparative structure, e.g., S_4) and explicit comparisons (mentioning both entities being compared, e.g., S_2).

We experiment with traditional machine learning methods and transformer-based models on Pixie. We use segment embeddings to demar-
cate the compared entities before fine-tuning the transformer-based models. We also compare our results with ED-GAT (Ma et al., 2020), a state-of-the-art model for preference classification. We find that transformer-based pretrained language models, fine-tuned on Pixie, achieve a higher F1-score (83.34%) than the state-of-the-art (F1-score 73.99%) or traditional machine learning models (F1-score 71.86%) trained on Pixie. Further error analysis of misclassifications reveals substantial differences between ED-GAT and transformer-based pretrained language models’ performance.

Current research on preference classification is lacking and far from practical use. Real world comparisons can present characteristics that complicate the task, such as indirect comparisons, implicit comparisons, and ambiguous statements. The low F1-score of the existing state-of-the-art and noticeable differences in misclassifications across different models call for a more thorough research effort on preference classification in text.

2 Related work

Comparative sentence structures have been a subject of syntactic and semantic theories (Bresnan, 1973; Stechow, 1984; van Rooij, 2011). Early studies in computational linguistics include syntactic and semantic handling of comparative constructions (Rayner and Banks, 1988, 1990), comparative structures in question answering (Ballard, 1988), using quantifiers to identify comparisons (Friedman, 1989), and semantic interpretation of comparatives (Staab and Hahn, 1997).

Jindal and Liu (2006) present a binary classification dataset containing comparative and non-comparative sentences. They present a classifier based on Class Sequential Rules (CSR) and leverage comparative keywords to identify comparative sentences. Ganapathibhotla and Liu (2008) extend this work by annotating comparative sentences with the preferred entity.

Kessler and Kuhn (2014) annotate comparative sentences by identifying comparison predicates, entities being compared, aspect of comparison, and comparison type (gradable or non-gradable). However, they focus on reviews of only one product type (digital cameras) to create their dataset. Hence, their dataset lacks diversity in topics.

Panchenko et al. (2019) create CompSent-19, a cross-domain dataset for comparative argument mining. They propose a gradient boosting model based on pretrained sentence embeddings to identify the preferred entity. Ma et al. (2020) propose a model called Entity-aware Dependency-based Deep Graph Attention Network (ED-GAT) that consists of a multihop graph attention network with dependency relations to identify the preferred entity. The ED-GAT model achieves a micro F1-score of 87.43% on the CompSent-19 dataset.

Previous work on preference classification has overlooked implicit and indirect comparisons common in user-generated text such as app reviews. Further, existing datasets are either too small with a few comparative sentences or have a skewed distribution. For example, Ganapathibhotla and Liu’s dataset contains only 837 comparative sentences, 84% of which have the first mentioned entity in the text as preferred. Only 15% of Kessler and Kuhn’s dataset constitutes comparative sentences. Only 27% of the sentences in CompSent-19 (Panchenko et al., 2019) contain a preference, 70% of which prefer the first mentioned entity in the sentence.

Further, existing datasets consider the order of the appearance of compared entities in a sentence to annotate the preferred entity. For instance, annotations for CompSent-19 (Panchenko et al., 2019) and Ganapathibhotla and Liu’s dataset are both determined based on the order of appearance of the entity in a sentence (i.e., is the first appearing entity in the sentence preferred or the second).

3 Method

We introduce the essential concepts below.

Comparative sentence: A sentence that contains information on similarity, dissimilarity, or preference between two entities.

Pixie includes (1) comparative sentences that lack comparative quantifiers, adjectives, or keywords, i.e., indirect comparisons, (2) implicit comparisons where only one of the compared entities is mentioned, and (3) explicit comparisons which mention both (including pronominal references).

Preferred entity: an entity that is chosen over another based on a stated or implied preference.

A preferred entity can be the CURRENT app (e.g., $S_{1p}$ in Table 2), OTHER app (e.g., $S_{2p}$ in Table 2), or NONE (i.e., ambiguous or no preference, e.g., $S_{4p}$ in Table 2) or where non-gradable comparatives (such as like, as . . . , and similar to) link the entities, e.g., $S_{3p}$ in Table 2).
Sentence Review
S1 This app is better than Discover’s app. Chase Mobile
S2 I prefer the BBC app. USA Today
S3 Just as good as Uber app. Lyft
S4 Makes me want to switch back to Pandora, but it’s just as bad. Spotify

Table 2: Example sentences showing preference.

3.1 Dataset

We selected 179 popular apps on Apple App Store and collected their reviews. After some preliminary investigation, we excluded widely mentioned brand names such as Google, Microsoft, and Facebook, because they often appear in broader contexts than as a product. We removed app names synonymous with or formed of common words, such as Box (cloud storage) and Line (communication app) for higher precision in extracting comparative sentences. We were left with 141 apps, which we manually grouped into 23 genres, including banking, airline, and communication. Apps in the same genre are direct competitors. For example, airline apps include Delta, American, and United.

We extracted sentences that mention a competitor from each review and labeled each extracted sentence for comparison and preferred entity. When identifying mentions, we included common aliases or abbreviations on our name list, e.g., Insta for Instagram, BA for Bank of America, and AA for American Airlines to improve recall. Focusing on mentions of competitors ensures that Pixie includes indirect comparisons because such sentences are more likely to contain comparisons.

The dataset was annotated in three phases. In Phase 1, the authors annotated a sample dataset of 300 sentences based on an initial set of definitions and resolved any disagreements via discussions. We repeated this process for three iterations and produced annotation instructions for Phase 2. In Phase 2, each author annotated an equal number of sentences, and the disagreements were resolved by the first author, producing 4,793 annotated sentences. The interrater agreement (Krippendorff alpha) was 0.74 and 0.82 between the two annotators for comparison and preferred entity, respectively. A portion of the dataset, ≈2,100 (~23.62%) sentences, had pronoun references that were resolved.

To ensure that the dataset can be used to train a general-purpose preference classification model, we mask app mentions in each sentence. With no masking, the model may learn to differentiate between classes based on what users prefer more (app A or app B) in our dataset. Masking app mentions ensures that the model learns comparative and preference revealing linguistic structures and semantics instead of simply learning to differentiate between preferred entities in an exhaustive list of compared entities. We defined two tags for masking, current_app for the apps being reviewed and other_app for the competitor apps. App mentions are identified using the competitor app list for apps referred to by name, and pronoun references are substituted manually. Treating pronoun references as an explicit reference to app mentions ensures consistent based on our definitions, i.e., all explicit comparisons have two mentioned entities being compared, while all implicit comparisons have one. A portion of the dataset, ≈2,100 (~23.62%) sentences, had pronoun references that were resolved. Table 4 shows sentences masked for app mentions.

For a quick sanity check, whether Pixie contains indirect comparative sentences, we examine how many of the sentences in Pixie contain a comparative word. For this, we combine the list of opinion words from (Hu and Liu, 2004) and the list of comparative cue words from (Panchenko et al., 2019). Only 3,781 sentences (42.5% of Pixie) contain a comparative or opinion word showing that most of the sentences in Pixie lack comparative cues (i.e., comparison and preferred entity, respectively. We obtained the Institutional Review Board (IRB) approval for this task.

Once we removed duplicate and noncomparative sentences, we were left with 8,890 comparative sentences annotated for comparison type (IMPLICIT or EXPLICIT) and preferred entity (CURRENT, OTHER, or NONE). Table 3 shows the distribution of labels for each class in Pixie.

Table 3: Pixie Dataset Distribution.

<table>
<thead>
<tr>
<th>Preferred Entity</th>
<th>Implicit</th>
<th>Explicit</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CURRENT</td>
<td>1,910</td>
<td>2,097</td>
<td>4,007</td>
</tr>
<tr>
<td>OTHER</td>
<td>2,199</td>
<td>1,069</td>
<td>3,268</td>
</tr>
<tr>
<td>NONE</td>
<td>758</td>
<td>857</td>
<td>1,615</td>
</tr>
<tr>
<td>Total</td>
<td>4,867</td>
<td>4,023</td>
<td>8,890</td>
</tr>
</tbody>
</table>

Table 4 shows sentences masked for app mentions.

For a quick sanity check, whether Pixie contains indirect comparative sentences, we examine how many of the sentences in Pixie contain a comparative word. For this, we combine the list of opinion words from (Hu and Liu, 2004) and the list of comparative cue words from (Panchenko et al., 2019). Only 3,781 sentences (42.5% of Pixie) contain a comparative or opinion word showing that most of the sentences in Pixie lack comparative cues (i.e.,
are indirect comparisons).

Unlike prior datasets on preference classification (Ganapathibhotla and Liu, 2008; Panchenko et al., 2019), Pixie does not consider the order of appearance of compared entities for annotations. Pixie also offers a more balanced dataset than the existing ones for the task. For explicit comparisons (when both entities are present), 1909 sentences (47.45%) prefer entity that appears first, 1257 (31.25%) sentences prefer entity that appears later, and 857 sentences (21.30%) reveal no or mixed preference. Implicit comparisons mention only one entity so the order of appearance is irrelevant.

Pixie is publicly available¹ and contains original and masked sentences.

3.2 Experiments

Among traditional machine learning approaches, we experiment with AdaBoost (Hastie et al., 2009), Random Forest (Breiman, 2001) and Support Vector Machine (SVM) (Chang and Lin, 2011). We use SBERT (SentenceBERT) (Reimers and Gurevych, 2019) to obtain sentence embeddings for each masked sentence.

For transformer-based language models, we fine-tune variations of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and XLNet (Yang et al., 2019). We experiment with BERT (Devlin et al., 2019), ALBERT (A Lite BERT) (Lan et al., 2019), and DeBERTa (Decoding-enhanced BERT with disentangled attention) (He et al., 2020). We fine-tune each model for 20 epochs using AdamW Optimizer with a learning rate of 5e-5 and a weight decay of 0.01. We use the train-test-validation split of 60-20-20.

We use segment embeddings to improve the performance of the transformer-based models. We assign different segment token ids to the competitor app (other_app) and the rest of the sentence to separate the entities being compared. We fine-tune pretrained models with segment embeddings along with token embeddings and attention masks.

To compare our results with ED-GAT, we convert the sentences in Pixie to follow the CompSent-19 format. Specifically, we add a token (THIS) for the current app in the front of each implicit sentence and map the labels CURRENT and OTHER to BETTER and WORSE, as applicable. NONE labels stay the same. We implemented ED-GAT with BERT embeddings and used eight GAT layers. We use the Hugging Face (Wolf et al., 2020) library for all transformer-based experiments.

To test the quality of Pixie, we run some cross-dataset experiments as well. We train a DeBERTa model on Pixie and test on CompSent-19 and vice-versa. Since the CompSent-19 dataset is highly skewed, we balanced both datasets to have the same train and test data split across all three classes via random oversampling with replacement. We keep all other model parameters and configurations the same and leverage the same number of samples for training and testing.

4 Results

Table 5 contains results for models trained and tested on Pixie. SVM achieves the highest weighted F1-score of 71.86% (among the traditional approaches), and DeBERTa (F1-score 83.34%), among transformer-based models.

Segment embeddings enhanced BERT and XLNet model’s performance in terms of weighted average F1-scores, but a slight decline for DeBERTa’s and ALBERT’s performance.

The NONE and CURRENT classes consistently achieve the lowest and the highest F1-scores, respectively, for all models. The NONE class was also the most ambiguous class to annotate manually. Recall for the NONE class is lower than precision for all models except ED-GAT. All transformer-based models achieve a higher recall than precision for the CURRENT class except for ALBERT (without segment embeddings) and ED-GAT.

ED-GAT (Ma et al., 2020) trained on Pixie achieves a weighted average F1-score of 73.99%, with the highest F1-score of 80.57% for the CURRENT class and lowest (51.54%) for NONE.

Upon further analysis, we found that most of the incorrect classifications in transformer-based models are for the NONE class (71.64%), whereas, for ED-GAT, only 8.77% of the misclassified sentences belong to the NONE class. ED-GAT yielded most misclassifications for the CURRENT class (55.27% of misclassified instances) while only 14.93% of misclassifications for the transformer-based models belong to the CURRENT class.

Table 6 shows the results for the cross-dataset experiments. The weighted average F1-score improves by 4.08% with plain vanilla fine-tuning and 6.30% with segment embeddings when trained on Pixie and tested on CompSent-19. While the accuracy improves by 5.11% for plain vanilla fine-
1. CNN should leave journalism to the pros at Fox news. 
2.  way better than Pandora by a long shot!!!!
3.  This is a great game just like Temple run

Table 4: Original and masked comparative sentences.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Model</th>
<th>CURRENT</th>
<th>NONE</th>
<th>OTHER</th>
<th>WEIGHTED AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>F1</td>
<td>Prec</td>
</tr>
<tr>
<td>Prior Work</td>
<td>ED-GAT</td>
<td>83.24</td>
<td>78.05</td>
<td>80.57</td>
<td>48.89</td>
</tr>
<tr>
<td>Traditional ML</td>
<td>AdaBoost</td>
<td>71.57</td>
<td>73.44</td>
<td>72.49</td>
<td>45.06</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>71.27</td>
<td>80.42</td>
<td>75.57</td>
<td>71.31</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>76.99</td>
<td>82.17</td>
<td>79.49</td>
<td>62.63</td>
</tr>
<tr>
<td>Transformer-Based</td>
<td>BERT</td>
<td>82.83</td>
<td>89.03</td>
<td>85.82</td>
<td>62.68</td>
</tr>
<tr>
<td></td>
<td>DeBERTa</td>
<td>88.34</td>
<td>90.65</td>
<td>89.48</td>
<td>64.56</td>
</tr>
<tr>
<td></td>
<td>ALBERT</td>
<td>87.83</td>
<td>87.28</td>
<td>87.55</td>
<td>61.37</td>
</tr>
<tr>
<td></td>
<td>XLNet</td>
<td>83.45</td>
<td>90.52</td>
<td>86.84</td>
<td>67.06</td>
</tr>
<tr>
<td>Transformer-Based with Segment Embeddings</td>
<td>BERT</td>
<td>83.43</td>
<td>88.53</td>
<td>85.90</td>
<td>66.67</td>
</tr>
<tr>
<td></td>
<td>DeBERT</td>
<td>88.31</td>
<td>91.40</td>
<td>89.83</td>
<td>64.34</td>
</tr>
<tr>
<td></td>
<td>ALBERT</td>
<td>86.26</td>
<td>87.66</td>
<td>86.95</td>
<td>65.92</td>
</tr>
<tr>
<td></td>
<td>XLNet</td>
<td>85.68</td>
<td>90.27</td>
<td>87.92</td>
<td>61.86</td>
</tr>
</tbody>
</table>
| Table 5: Results (in %) for preference classification on Pixie. Bold indicates highest F1-scores for each category.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Fine-tuning</th>
<th>Testing</th>
<th>Prec</th>
<th>Recall</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain vanilla</td>
<td>CompSent-19</td>
<td>Pixie</td>
<td>65.46</td>
<td>59.89</td>
<td>59.23</td>
<td>59.89</td>
</tr>
<tr>
<td>Plain vanilla</td>
<td>Pixie</td>
<td>CompSent-19</td>
<td>65.19</td>
<td>65.00</td>
<td>63.31</td>
<td>65.00</td>
</tr>
<tr>
<td>With segment embeddings</td>
<td>CompSent-19</td>
<td>Pixie</td>
<td>67.84</td>
<td>59.44</td>
<td>57.70</td>
<td>59.44</td>
</tr>
<tr>
<td>With segment embeddings</td>
<td>Pixie</td>
<td>CompSent-19</td>
<td>66.07</td>
<td>65.72</td>
<td>64.00</td>
<td>65.72</td>
</tr>
</tbody>
</table>

Table 6: Results for cross-dataset experiments. The values are in %.

5 Conclusion

Masking compared entities ensure that Pixie can be used to train a general-purpose preference classification model. Additional analysis is needed to claim the domain generality of our dataset—that is, whether a model trained on Pixie can identify the preferred entity in texts from other domains such as scientific papers and news. Comparative sentences in Pixie are limited to user-generated text and may not generalize well over more formal texts.

Both BERT and XLNet show improvements with segment embeddings, suggesting that the demarcation of the other app helps the model identify the preferred entity. The traditional machine learning models perform worst and the transformer-based pretrained models fine-tuned on Pixie achieve a substantially better performance than the state-of-the-art approaches for preference classification.

Identifying preferences in user reviews can aid developers in understanding user expectations about mobile apps. Users often express their likes and dislikes about an app or feature by comparing it with alternative apps and features. Understanding user preferences can be particularly valuable in enhancing the functionality as well as security and privacy features of apps. A user’s preferences regarding apps would depend not only on how well the app is constructed relative to its competitors but also on how easily the app is used by end-users. For example, security concerns may be signaled by descriptions of steps to access sensitive financial or medical data (Guo and Singh, 2020) expressed in association with comparisons. A follow-on direction is to extract and prioritize user expectations by identifying the specific features of an app of greatest influence on the indirect or direct comparisons in a review.
Acknowledgments

This research is supported by the US Department of Defense under the Science of Security Lablet (SoSL) grant to NC State University. We thank the ACL reviewers for their helpful comments.

References


