An Approach to Identifying Aspects of Positive Pilot Behavior within the Aviation Safety Reporting System

1st Bryan Matthews
KBR, Inc.
NASA Ames Research Center
Moffett Field, USA
bryan.l.matthews@nasa.gov

2nd Immanuel Barshi
NASA
NASA Ames Research Center
Moffett Field, USA
immanuel.barshi@nasa.gov

2nd Jolene Feldman
NASA
NASA Ames Research Center
Moffett Field, USA
jolene.m.feldman@nasa.gov

***Abstract*—The National Airspace System (NAS) is constantly evolving as air traffic continues to ramp up to pre-pandemic numbers and projected to grow to unprecedented levels in the coming years. As well as increasing demand to the current system, Emerging operations such as Unmanned Autonomous Systems are also expected to add to complexity to the airspace. To address these issues, the industry and government agencies supporting the NAS will need to rely upon additional automation and new technologies to address future operational requirements, while continuing to be a world-leading safe transportation system. As these new technologies are implemented, the system continues to rely on human pilots and controllers in the loop to monitor the system and intervene in situations the automation cannot handle. The goal of proactively addressing safety is of foremost concern to ensure passenger well-being. The industry has implemented various Safety Management Systems to identify safety risks and proactively address them before they result in a serious incident or accident. One such program is NASA's Aviation Safety Reporting System (ASRS). ASRS is a long-established system where pilots and controllers voluntarily and anonymously report safety incidents they experienced and observed during line operations by providing rich text narratives describing the events, the environment, and conditions leading to the safety event of concern. These narratives provide insight and context around events of interest and can be used to identify emerging problems. These reports can also trigger investigations within Flight Operational Quality Assurance or Flight Data Monitoring programs. However, this process typically focuses on the adverse events and the unsafe aspects of the operations surrounding the reported or detected events. This perspective of investigating factors that went wrong around an adverse event is commonly referred to as Safety I. Alternatively, characterizing successful actions that operators perform every day under varying conditions that keep the system within safe operating bounds is a concept referred to as Safety II. The benefit of the Safety II view is that its scope is much larger than that of Safety I since a vast majority of the operations result in successful flights. Many of the successful techniques used to manage operational threats are not documented in standard operating procedures or taught during training. They are typically acquired over time by working with experienced pilots during line operations or in many cases after experiencing a problem for the first time and reacting to it in situ, drawing from years of experience to manage the threat. In an attempt to quantify these positive actions, we are proposing an approach to extract key behaviors within ASRS reports that can support the Safety II concept. Our analysis assumes that ASRS reports contain some descriptions of corrective actions that operators performed to prevent a situation from becoming an accident. Leveraging recent advances in Natural Language Process (NLP) modeling, we have developed an approach to extract positive sentiment from reports, embed these positive statements in a vector space where they can be numerically analyzed, and clustering these statements into similar contextual categories. From these contextualized categories we can attempt to summarize and distill aspects of the positive behavior. The goal is to identify categories of behavior that describe consistent operator techniques that supports the Safety II concept. With this information, airlines may enable learning from these positive actions, or address procedures that need to be changed. These insights can provide a lens into what is "going right" in the operations that may otherwise not be known widely within the community. It is envisioned that this approach can be extended to other narrative programs such as Line Operation Safety Audit or Learning Improvement Team reports where similar observed behavior can be analyzed to extract positive actions and inform the overall operations.**

***Index Terms*—NLP, ASRS, Aviation Safety, Human Factors, Text Mining**

I. Introduction

A. Background

The main objectives for pilots and controllers is to transport passengers and cargo safely to their destination while attempting to stay as close to the flight schedule as possible. This objective influences these operators' decisions and guides the actions that they take throughout the course of a flight. However pressures such as weather, high traffic demand, latent procedure complexity, disruptive passengers, or mechanical and software issues require positive actions to respond to these pressures to maintain safety. US Domestic carriers are required to maintain a Safety Management System (SMS) program by regulation [1] where airlines monitor and track safety events in Flight Operational Quality Assurance (FOQA), and Aviation Safety Action Program (ASAP) data. These data sources contain examples of both safety events as well as corrective actions that bring the aircraft back to a safe state. The approach

Funding supported by NASA's System-Wide Safety Project

for many SMS programs is to focus on the occurrences and conditions where these safety events manifest themselves and attempt to devise a safety mitigation strategy to reduce the risk of these events in the future. This concept of analyzing and mitigating adverse events is commonly referred to as the Safety I [2] concept. However, due to the resilient behavior of operators these events caused by pressures are typically resolved to prevent a more serious incident or accident from occurring. The analysis of these positive behaviors supports the Safety II concept. Some risk mitigations are built into the operations such as checklists and procedures that help create safety barriers in the system. At the same time, there are many actions that operators take that are not codified in procedures or checklists. These actions rely upon prior experience to respond to pressures that emerge during a flight operation. These actions can take the form of tactical responses such as countermeasures to pressures or strategically deployed modifications that are positive deviations from the planned operation [3]. We are proposing a process to leverage rich text narratives to extract fundamental actions that pilots take that result in success, and to identify instances where such actions are consistently being taken to maintain safety. Identifying these actions can be used proactively to reinforce pilot proficiencies and to impart knowledge to low-experience pilots who have not yet gained these skills and techniques. Furthermore, understanding what actions are being performed by humans will help inform design of future automation requirements needed to make the system more robust and resilient to safety pressures in the NAS.

B. Aviation Safety Reporting System

NASA's ASRS [4] database is the world's largest repository of safety reports. It contains voluntary text narratives from line operators such as pilots, controllers, flight attendants, and other members of the aviation community. Many of the ASAP reports filed at individual airlines are also uploaded into the ASRS database where they are de-identified to preserve the anonymity of the airline and the individual who filed the report. The database is comprised of nearly 2 million narratives which describe safety events that were experienced and the actions taken to resolve them. These positive actions support the Safety II concept and have not been fully explored with machine learning (ML) in combination with natural language processing (NLP) techniques within this context. Positive actions found within ASRS may potentially highlight resilient behaviors that are being performed each day and are a critical contribution to the safety of the NAS.

II. Methods

A. Sentiment Analysis using RoBERTA

Early work in sentiment analysis relied upon small labeled or crowd sourced data sets to build ML model for predictions. Pang et al. [5] and Turney et al. [6] both led the way in sentiment analysis in 2002. They relied upon small datasets on the order of 1,000 movie reviews to build their models. Pang utilized Naive Bayes classifier whereas Turney used a semi-supervised approach that leveraged a Pointwise Mutual Information and Information Retrieval (PMI-IR) technique. Pang and Lee in 2005 [7] introduced a balanced data set consisting of approximately 10,000 movie reviews[1]. In 2011 Socher et al. [8] was the first to construct a recurrent neural network (RNN) based model to address the sentiment classification problem utilizing this curated movie review database. However, there are some drawbacks from using models based on this movie review data. One is that it contains a unique lexicon that is used to describe the overall sentiment as it pertains to a film, which does not always transfer to other domains such as aviation. Another disadvantage to using the reviews is that each review is assigned a sentiment based on the rating and the reviews are typically over 30 sentences long. Not all the sentences within a review are all positive or all negative to match the review rating and therefore the labels for each review may not capture the correct sentiment of each sentence. Barbier et al. [9] utilized a data set of an approximately 60M Twitter posts. Each post was labeled by inferring the sentiment based on the types of emojis used in the text. Although these data do not specifically discuss aviation topics, it is orders of magnitude larger than the movie database and contains discussions over a variety of domain topics. Additionally, each post is typically only a couple of sentences long and the sentiment, therefore, is less likely to vary across the sentences within a post with respect to the emoji sentiment label. Although there is potential for ironic or sarcastic use of the emojis in the tweets, the algorithm does not specifically address this potential issue and therefore how this effects the model is unknown. The model was built using the Robustly Optimized BERT Pretraining Approach (RoBERTA) [10], which is a variant on the Bidirectional Encoder Representations from Transformers (BERT) developed by Devlin et al. [11]. RoBERTA differs from BERT in two ways: (1) it removes the next sentence prediction optimization task and (2) introduces a dynamic token masking between training epochs. These advances in RoBERTA have been shown to consistently improve the model performance over BERT.

B. Primitive Sentence Structure

To help distill the clear action in a sentence and remove descriptive modifiers the sentence structure needs to be reduced to a primitive form. Using Python's NLTK Parts of Speech (POS) tagging, the sentences in our dataset were parsed in the following order: personal pronouns (PRP), verb (VB), and noun (NN). The justification for this ordering of the POS is so that the primitive sentence would read as "someone did something". A few commonly used aviation abbreviations were hard coded as PRPs. These were: tower, clt (controller), PF (pilot flying), PM (pilot monitoring). It was important to capture these key PRP subjects so if there were any clear consistent behaviors they would be tied to these performers. If there were multiple nouns at the end of the sentence all

[1]Available at http://www.cs.cornell.edu/People/pabo/movie-review-data as scale dataset v1.0.

were included, providing flexibility for the sentence to have multiple outcomes resulting from the given action.

C. Text Embedding

Text embedding is a useful technique to translate words from the natural language into an embedding vector space. Having the text translated into this space allows for advanced numerical processing techniques to be utilized. Word2Vec [12] is an algorithm that builds such a model in a multi-dimensional embedding space. It uses a neural network to learn word associations between a target word and other words within a given window. The algorithm is unsupervised, meaning that no labels are provided by a subject matter expert to train the model. The model's task is to predict the surrounding words based on the target word. This approach is referred to as the skip-gram method. The neural network architecture uses a single fully connected layer to map the input word space to a user defined latent space dimension and then predicts the output words from that latent space. Once the model is trained, a word can be directly mapped into the latent embedding space where words with similar meaning are highly correlated using a distance metric such as cosine similarity. For our purposes we chose a 300 dimension embedding vector based on Mikolov's prior work [12].

To prevent highly occurring uninformative terms from dominating the word embedding space a technique referred to as Term Frequency-Inverse Document Frequency (TF-IDF) [13] was used to weight the Word2Vec embedding vector. The weighting de-emphasizes words that appear across a majority of the reports and boosts terms that occur more frequently within a single reports. For each sentence, Word2Vec maps each word to an embedding vector and the TF-IDF weighted average is representative of the sentence's overall embedding vector.

D. Clustering with HDBSCAN

Hierarchical Density Based Clustering (HDBSCAN) [14] is a clustering based approach that has qualities that are amenable to the objectives for our analysis. The first is that density based clustering has the ability to identify points that fall within a background cluster. These points do not have a clear cluster membership and can be considered noisy data points. We assume that not all sentences can find a corresponding cluster, allowing for those sentences to be grouped in this catch-all cluster. It is assumed that because such sentences are not like other sentences, they do not have the consistent behavior we are looking for and can be ignored. The next benefit of HDBSCAN is that it is an improvement over DBSCAN (Density Based Clustering) in that HDBSCAN explores varying epsilon values for the cluster density cutoff threshold and automatically selects the epsilon value that provides the best stability across the clusters, which removes the need for hyperparameter tuning of this variable.

III. Methodology

We analyzed an ASRS reports database that contained 239,657 individually filed reports covering 221,551 unique

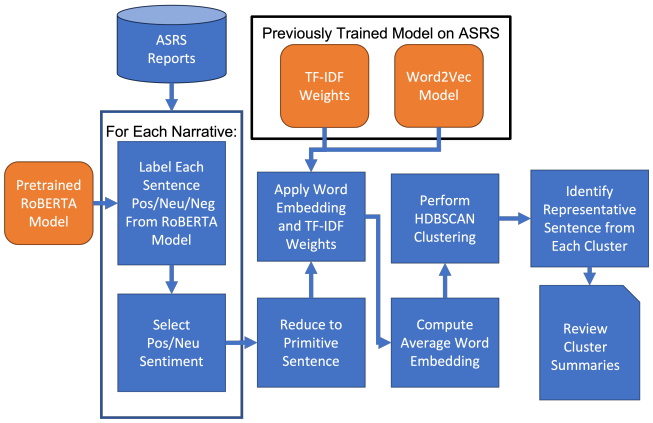

Fig. 1. Data Processing Pipeline

events (events may have reports filed by both pilots or by a pilot and a controller) submitted between the years 1988 and 2022. The database contains both reports filed by general aviation pilots as well as commercial operations. The first step in the process is to apply common NLP techniques such as removing stop-words and stemming from the ASRS reports. This was implemented with Python's NLTK package's [15] using the 'english' stopword list. The next step is to compute TF-IDF weights and train the Word2Vec model. The pretrained RoBERTA sentiment model found on Huggingface's open sourced model repository[2] was downloaded and applied to each sentence in the ASRS reports. Negative sentences were dropped leaving 368,206 sentences (29,896 positive and 338,310 neutral). After identifying the positive and neutral sentence each sentence was broken down into the primitive sentence structure discussed in section II-B and mapped to the Word2Vec 300 dimensional average embedding space. HDBSCAN clustering was performed resulting in 3,274 clusters identified as well as the background cluster. The background cluster contained 186,819 sentences (50.7% of the total number of sentences). The average cluster size was on the order of 50 sentences with a few in the hundreds. A minimum cluster size of 15 sentences was chosen based on the default used in the HDBSCAN documentation[3].

To help with understanding the clusters a summarizing technique that looks at the most representative sentence was implemented. This was done by taking the original sentences in each cluster, mapping them to the average embedding space, similar to what was done with the primitive sentences, and computing the centroid of the cluster in the embedding space. The sentence that is the closest to this point is considered the medoid or representative center of the cluster. Figure 1 shows the overall pipeline of the methodology.

IV. Analysis And Discussion

Due to the large number of clusters produced, it was infeasible to examine every cluster by hand and summarize

[2]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment
[3]https://readthedocs.org/projects/hdbscan/downloads/pdf/latest/

its positive behavior. However, a subject matter expert (SME) reviewed a select number of clusters and assigned a summary to each of the sentence clusters. The majority of the SME's summaries aligned well with the representative sentence with each of the samples of clusters reviewed. Although many of the sentences within a cluster described consistent actions, there were some sentences that contained a relevant word to that cluster, that word was used in a different context. The SME noted that some clusters appeared to be more affected by this phenomena than others. This may be an artifact of the clustering process and could be addressed in the future with better hyperparameter tuning or by performing cluster quality analysis to provide a way to rank order clusters with more consistent sentences. The summaries were examined and manually grouped into the following five different high-level categories of actions:

1) Addressing Automation Failure (Table I).
2) Managing Aircraft State (Table II).
3) Assessing Aircraft (Table III).
4) Diagnosing and Responding to Sensor/System Failures (Table IV).
5) Managing Latent Procedure/Air Space Complexity (Table V).

Three of the high-level cluster categories $\{1,2,4\}$ are actions where pilots were required to take control when automation limits are exceeded or to diagnose a failure in the system. These categories of events fall within failures of the existing automation. Moreover, these highlight deficiencies in existing systems that original equipment manufacturers need to address in order to maintain the current level of safety as more automation is introduced in the operations. If these systems can be improved to an acceptable 10^{-9} level of failure rates, then there is potential to utilize these systems without relying upon the pilots to resolve such system problems.

The remaining two high-level cluster categories $\{3$ & $5\}$ demonstrate complex reasoning and sensing from the pilots. Some of the pilots' actions are performing assessments based on visual observations. Other aspects of these actions require making judgement calls or anticipating potential threats when information is requested or presented to the pilots. These complex decision making processes can require bringing together multiple sources of information and determining the best course of action to correctly respond while meeting the objective of staying as close to the schedule as possible and maintaining a safe aircraft. There are currently no capabilities on the aircraft that can replicate or automate these human capabilities. These tasks would require additional sensors on the aircraft as well as complex decision making algorithms. These adaptive algorithms would need to be able to make similar judgement assessments on the aircraft's state of health for the system to be fully automated.

TABLE I
CLUSTER SUMMARY TABLE

Addressing Automation Failure		
Cluster	Summary	Representative Sentence
4	Switching to manual flying.	I attempted to disengage autopilot (ap) and manually fly the turn and climb," the ap would not initially disengage.
24	Taking control with failed flight director.	During the time I was doing said things above (hot app, deice checklist, and fd) the crew called for taxi.

TABLE II
CLUSTER SUMMARY TABLE

Managing Aircraft State		
Cluster	Summary	Representative Sentence
150	Managing engine EGT exceedances.	We turned off the auto throttles and reduced climb rate while reducing power on left engine to get EGT back in normal range.
215	Recognizing aircraft speed deviations from ref.	We never had a high sink rate or airspeed, and we were right on vref in the turn.
239	Adjusting thrust to manage fight path.	However once we landed my brain revered to muscle memory and after touchdown, I pulled both thrust reverser triggers.
242	Monitoring deviations from glideslope.	We continued uneventfully", staying one dot high on the ILS.
305	Using configuration changes to adjust the flight profile.	We slowed and put out drag.
338	Using the VASI to verify vertical path during final approach.	I was following the locilizer (LOC approach) and using the VASI.
530	Correcting vertical path issues.	I immediately went to the autopilot disconnect switch on the yoke while trying to arrest the descent with elevator back pressure
532	Managing windshear events.	While on final, we received another windshear warning between 1000-1500 feet.

V. SUMMARY AND FUTURE WORK

The proposed methodology demonstrates how positive pilot behavior can be captured from ASRS reports and yield informative consistent positive pilot behaviors. This proof of concept shows that utilizing transfer learning of a pretrained sentiment labeling model can help to extract important positive aspects of narratives despite the fact that the reports were primarily written in response to a negative event. The methodology demonstrates the capability and potential to identify those positive actions which have previously gone unexplored within this rich data set. Furthermore, the positive behaviors identified highlight the different levels of automation that require further advancements before future systems can achieve more autonomy. These areas where advancements are needed are: (1) improving upon current automation robustness to prevent pilots from having to resolve automation failures and (2) designing new automation capabilities that capture both the complex sensing capabilities that pilots leverage as

TABLE III
CLUSTER SUMMARY TABLE

Assessing Aircraft		
Cluster	Summary	Representative Sentence
5	Assessing damage after an event.	We were all pleasantly surprised to find no visible damage and everything intact
15	Examining plane.	Also, I told the inspector that a pilot from cirrus was coming with the mechanic from cirrus so he can examine and determine if a flight was possible.
46	Noticing ice on the wings and control surfaces.	I asked him to point the nozzle at the aircraft surfaces, and the frost.
142	Assessing the aircraft's condition.	He then called them himself to apprise them of our condition.

TABLE IV
CLUSTER SUMMARY TABLE

Diagnosing and Responding to Sensor/System Failures		
Cluster	Summary	Representative Sentence
7	Switching to directional gyro due to magnetic interference.	We then realigned our headings to match the runway and elected to depart in DG mode as per our procedures due to the magnetic interference.
8	Noticing Internal Reference System failure.	I suspect they may need to replace the IRS
12	Trouble-shooting generators.	I began that portion of the checklist which directs you to turn off the DC and then the AC generators.
14	Noticed braking problems.	I think the ramp was the last area to be treated for the coming precipitation and it was nearly an hour after we had reported NIL braking action.
28	Responding to fuel imbalances.	I noticed a slight imbalance of about 300 pounds so I turned the center tank pump switch off and left the crossfeed open.
35	Trouble shooting faulty warning horns.	So I pulled the number 2 (failed engine) power lever aft to idle in order to activate the gear warning horn (sonalert).
50	Solving complications arising from CPDLC errors.	As we approached a line of severe weather we requested a deviation via CPDLC.

TABLE V
CLUSTER SUMMARY TABLE

Manage Latent Procedure/Air Space Complexity		
Cluster	Summary	Representative Sentence
9	Anticipating an unusual procedure at WENTZ.	We were cleared for takeoff and shortly after while i was running climb checks and switching to departure, I noticed that we were going through 1,500 ft (I think we were at 1,800-1,900 ft) before WENTZ.
21	Managing spacing with proceeding airbus.	We were told by approach that our spacing looked good with the airbus ahead of us.
61	Maneuvering the aircraft to avoid threats.	They both had at this point started into a TCAS maneuver.
83	Coordinating with supervisory structures over ACARS to manage irregular operations.	We were also getting ATIS and in-range information on ACARS.
171	Integrating information from the HUD into flight path management.	At approximately 1000 feet AGL I disconnected autopilot/auto-throttles (or so I thought) to continue the approach hand flying using the HUD.
337	Addressing threats posed by thunderstorms.	I stated in a loud and clr voice", that this would take me into a known tstm.
520	Prioritizing tasks during abnormal situations.	I was intent on flying the acft while f/o dealt with prob.
588	Deciding to divert to an alternate.	We had augmented crew with two other pilots.

well as the advanced decision making that humans have the ability to perform.

Moreover, there is still a potential for continued advancement with this methodology in our future work. All aspects of the methodology's process have the potential for improvement. Exploring other sentiment models or fine tuning the existing sentiment model with more domain specific sentiment data could offer some advantages. Refining the parts of speech assignments and primitive sentence structure has the potential to yield more clear actions and outcomes. Improving the clustering process and organizing the clusters is another area to explore. Some sentences containing rare key words were clustered together, however, the context around the key word was not consistent across all sentences. This may be due to the TF-IDF weighting assigning too much emphasis on that word due to its rare occurrence. In future work, we will investigate if this can be addressed. Capturing the cluster quality or improving the sentence embedding vector may be another way to address this issue. Another aspect of clustering is to address the large number of clusters being formed by better organizing them into aggregate themes so that they can be explored more efficiently by the analysts and ultimately assist stakeholders in making decisions. Identifying relevant and non-relevant actions found by this methodology is still done by hand. Finding ways to automatically filter out non-relevant actions will help with adoption and bring the most relevant behaviors to the attention of the stakeholders who can then utilize this valuable information.

ACKNOWLEDGMENT

We would like to thank Captain Barth Baron for his subject matter expertise as well as NASA's Human Contribution to Safety team for their valuable feedback and comments. And we would like to acknowledge the System-Wide Safety project under NASA's Aeronautics Mission Directorate's Airspace Operations and Safety Program for funding this work.

REFERENCES

[1] U.S. Dept. of Transportation Federal Aviation Administration, "Safety management system." https://www.faa.gov/documentLibrary/media/Order/Order_8000.369C.pdf, 2020.

[2] E. Hollnagel, R. L. Wears, and J. Braithwaite, "From safety-i to safety-ii: a white paper," *The resilient health care net: published simultaneously by the University of Southern Denmark, University of Florida, USA, and Macquarie University, Australia*, 2015.

[3] M. Stewart and B. Matthews, "Resilient strategies in commercial aviation," in *22nd International Symposium on Aviation Psychology*, Proceedings, (Rochester, NY), 2023.

[4] NASA Aviation Safety Reporting System, "Asrs database online." https://asrs.arc.nasa.gov/search/database.html, 2023.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 79–86, Association for Computational Linguistics, July 2002.

[6] P. D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," 2002.

[7] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the ACL*, 2005.

[8] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, (Edinburgh, Scotland, UK.), pp. 151–161, Association for Computational Linguistics, July 2011.

[9] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 1644–1650, Association for Computational Linguistics, Nov. 2020.

[10] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp. 502–518, 2017.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[13] A. Rajaraman and J. D. Ullman, *Data Mining*, p. 1–17. Cambridge University Press, 2011.

[14] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Advances in Knowledge Discovery and Data Mining* (J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, eds.), (Berlin, Heidelberg), pp. 160–172, Springer Berlin Heidelberg, 2013.

[15] E. Loper and S. Bird, "NLTK: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, (Philadelphia, Pennsylvania, USA), pp. 63–70, Association for Computational Linguistics, July 2002.