Colton, David, and Hofmann, Markus (2019) Sampling Techniques to Overcome Class Imbalance in a Cyberbullying Context. Journal of Computer-Assisted Linguistic Research issue: 21-40.

Sampling Techniques to Overcome Class Imbalance in a Cyberbullying Context

David Colton*[1], Markus Hofmann[2]
[1] IBM, Ireland
[2] Technological University Dublin, Ireland
*Corresponding author: DavidColton@ie.ibm.com

Received: 15 December 2018 / Accepted: 26 February 2019 / Published: 15 July 2019

Abstract
The majority of datasets suffer from class imbalance where samples of a dominant class significantly outnumber the samples available for the minority class that is to be detected. Prediction and classification machine learning models work best when there are roughly equal numbers of each class type. This paper explores sampling techniques that can be used to overcome this class imbalance problem in a cyberbullying context. A newly classified cyberbullying dataset, including detailed descriptions of the criteria used in its classification, was used to examine the feasibility of applying text mining techniques, to automate the detection of cyberbullying text when the dataset shows a significant class imbalance between the positive, cyberbullying, sample and the negative, not cyberbullying, samples. In this paper, we will investigate if oversampling the minority positive class or undersampling the majority negative class affects the performance of a prediction model. A compromise solution where the positive class is partially oversampled, and the negative class is partially undersampled is also examined. Although not strictly a class imbalance solution, sampling using the most frequently observed features was also explored.

Keywords: text mining, class imbalance, cyberbullying, sampling, classification.

1. Introduction

According to Nahar, Li and Pang (2013), bullying has moved out of the school yard and is now causing concern online as cyberbullying. Dadvar, et al. (2012) highlight that in this age of digital communications you can have hundreds of virtual friends having never met them face-to-face. Kontostathis, Reynolds, et al. (2013) suggest that the internet and social media applications are being used, particularly by children and teenagers, as a new way to bully, to cyberbully.

The U.S. Department of Health and Human Services (2018) (DOH) highlights that bullying exists where there is either a perceived or actual imbalance of power and there is repeated

DOI: https://doi.org/10.4995/jclr.2019.11112
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

aggressive behaviour. These behaviours could take many forms including verbal, for example teasing or name calling, social, including spreading rumours and exclusionary acts and threatened or actual physical violence. The DOH describes cyberbullying as bullying using communication tools such as instant messaging, chat sites and social networks using smart phones, computers or tablets. The DOH also highlights that cyberbullying cannot be easily turned off and can happen twenty four hours a day seven days a week. To further exasperate the situation the bullying messages or texts can rapidly spread to a large on-line community. Also, the bully can be anonymous and difficult to trace and completely purging the internet of the offending text or image is next to impossible. The persistent nature of the world wide web and functionality provided by social media sites means that these abusive posts can quickly spread amongst a group and can subsequently be accessed again and again by both the victim and the perpetrator. The implications of this is that the offending text or image can continuously reappear causing distress to the victim again and again. The effect of the cyberbullying can be traumatic on the victim: leading to trouble sleeping, withdrawing from society, stress and more troubling mental health problems like anxiety, depression and suicidal thoughts.

The affects of cyberbullying on its victims can sometimes have tragic consequences. News headlines like "Third suicide in weeks linked to cyberbullying" (Cionnaith 2012), "Cyberbullies claimed lives of Five teens" (Riegel 2013) and "Hanna Smith suicide fuels calls for action on Ask.fm cyberbullying" (Smith-Spark 2013) are now, unfortunately, becoming an all too depressingly frequent occurrence. However, behind these eye-catching headlines the daily torment and abuse suffered by the victims of cyberbullying are taking a dire psychological and emotional toll. The negative affects of posts that contain cyberbullying of a personal or sensitive nature can be internalised by children and young adults leading to significant and emotional psychological suffering (Dinakar, Reichart and Lieberman 2011). Apart from suicidal thoughts (Xu, Zhu and Bellmore 2012), the affects of cyberbullying can include loneliness, anxiety, low self-worth and signs of depression. Other signs described are intra-personal problems, school absence or violence and physical complaints (Xu, Jun, et al. 2012).

It is clear that cyberbullying can negatively impact the quality of a teenagers life in many different ways. The victim of bullying can suffer physical stress and a range of emotional feelings including humiliation, isolation, powerlessness, feeling overwhelmed, depressed and even suicidal thoughts. These are feelings that young people may not be emotionally mature enough to handle. This emotional turmoil can lead to a loss of appetite and an inability to sleep which can cause other more serious health problems. By detecting and identifying the bully, intervention may be possible.

Undertaking any machine learning research in the area of cyberbullying is difficult, both because of the lack of suitably classified datasets but also because any datasets that are available mostly suffer from a class imbalance where the majority (not bullying) class usually significantly outnumber the minority (bullying) class.

The contributions of this paper are twofold. First, following a review of related work, a new dataset for use in cyberbullying research, and the criteria used in its classification is presented.

Next, we will show that, when presented with an imbalanced dataset, an improved model performance can be achieved using undersampling of the majority class, oversampling of the minority class or a hybrid approach that uses a combination of both.

2. Related Work

Throughout the literature review certain obstacles to the automated detection of cyberbullying text appeared multiple times. This next section will give a brief overview of these issues and, where offered, some of the suggested solutions. The topics include the availability of suitable datasets, classification or labelling, and class imbalance.

2.1. Data and Classification

Nearly every major paper reviewed expressed frustration at the lack of a standard labelled dataset that could be used in the research of cyberbullying detection. Yin, et al. (2009) and Nahar, Li and Pang (2013) both used the MySpace, Kongragate and Slashdot datasets from Fundación Barcelona Media datasets provided for the CAW 2.0 Workshop (FBM 2009). Nahar, Li and Pang also used the manually labelled dataset from Yin, et al. as their ground truth. Dinakar, Reichart and Lieberman (2011) used comments from YouTube videos as their data, treating each comment as a stand-alone comment. Dadvar, Trieschnigg, et al. (2013) also used YouTube video comments but Xu, Zhu and Bellmore (2012) and Xu, Jun, et al. (2012) used Twitter tweets as their dataset. Kontostathis, Reynolds, et al. (2013) and Reynolds, Kontostathis and Edwards (2011) both used data that was scraped from the www.formspring.me website.

Once the data had been acquired, the next issue was the classification or labelling of the data to identify those posts which were considered to contain cyberbullying content and posts that did not. In most cases, the data was manually classified by annotators known to the authors using a simple binary label (Dadvar, Ordelman, et al. 2012; Dadvar, Trieschnigg, et al. 2013; Dinakar, Reichart and Lieberman 2011; Yin, et al. 2009). The Mechanical Turk service offered by Amazon was also used on occasions (Kontostathis, Reynolds, et al. 2013; Reynolds, Kontostathis and Edwards 2011). Xu, Zhu and Bellmore (2012) and Xu, Jun, et al. (2012) used Twitter tweets for their dataset and, rather than classifying by hand, an enriched dataset was created by automatically filtering on tweets that contained the "bully", "bullying" and "bullied" keywords. In Chen, et al. (2012) a semi-automated process was used where words were automatically identified as profane using a dictionary and weighted according to their offensiveness.

2.2. Class Imbalance

It was also recognised that there existed a class imbalance between the positive bullying class, and the negative class (Dadvar, Ordelman, et al. 2012; Chen, et al. 2012). Reynolds, Kontostathis and Edwards (2011) found that only 7.2% of their training dataset was given a positive classification. As a result of this imbalance, it was observed that the chosen learner could achieve accuracy figures of over 90% by ignoring the cyberbullying cases and by just labelling everything as not containing cyberbullying. An approach to address this issue is replicating the positive classes (Yin, et al. 2009; Reynolds, Kontostathis and Edwards 2011).

Weiss, McCarthy and Zabar (2007) suggests that most classifiers are designed to maximise accuracy implying that when used to label a highly imbalanced dataset the more frequently occurring class will be predicted. However, when labelling such a highly skewed dataset it is usually the case that it is the less frequently occurring class that is of interest, for example in medical diagnostics and fraud detection. Three methods were evaluated to handle class imbalance. The first method uses Random Undersampling (RUS) of the majority class. The second uses Random Oversampling (ROS) of the minority class. The third method is a cost-based method where the cost of misclassification is built into the learner. Under and oversampling are not without their disadvantages. Undersampling runs the risk of discarding potentially useful data whilst oversampling, besides increasing the learning time required, can lead to overfitting, or the generation of a rule specifically for the replicated data. In their experiments a number of datasets were used with the C5.0 cost sensitive decision tree learner (RuleQuest Research), an enhanced version of the C4.5 learner (Quinlan 1993). For the experiments, the unbalanced datasets were submitted to the C5.0 learner with a variety of costs for misclassification. Each dataset was then rebalanced using both over and undersampling to replicate a ratio that mimicked the cost of misclassification and submitted to the C5.0 learner without a cost. Though the overall result was inconclusive, by measuring the total cost of misclassification at the various costs / ratios it was found that for larger datasets, greater than 10,000 examples, the cost sensitive algorithm outperformed both over and undersampling methods. Oversampling was found to perform the best on smaller datasets, in the experiments this was on datasets of less than 250 examples.

Kubat and Matwin (1997) refers to the problem when one dataset is significantly larger than the other as "Addressing the Curse of Imbalanced Training Sets". They identify that it is usually the minority class that is of interest and point out that a classifier that achieves 99.8% accuracy where the minority positive class only consists of 0.2% of the samples is, in fact, of no use at all if it the presence of the samples of interest are completely ignored. In addition, to standard classifier performance measures, for example accuracy, precision, recall and F-Measure, Kubat and Matwin describe another measure, called the g-performance, that uses the geometric mean of the accuracies measured separately on each class as:

Equation 1. G-Performance Measure.

$$g = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}$$

where:

1. **True Positive (TP)**: correctly identified i.e. bullying identified as bullying
2. **False Positive (FP)**: incorrectly identified i.e. not bullying identified as bullying
3. **False Negative (FN)**: incorrectly rejected i.e. bullying identified as not bullying
4. **True Negative (TN)**: correctly rejected i.e. not bullying identified as not bullying

The goal of this measure is to maximise the accuracies of both classes but at the same time keeping them balanced such that a poor value for either the positive, or negative class, will give an overall poor performance for the classifier.

There are other papers that suggest alternative approaches to tackle class imbalances. Chawla (2002) uses a combination of undersampling of the majority class and oversampling of the minority class using their novel algorithm SMOTE, Synthetic Minority Over-sampling Technique, to create synthetic minority class examples. Cardie (1997) presents two case-based learning frameworks in a natural language environment that uses information gained from analysing a baseline case to determine the appropriate class weighting. In Chan and Stolfo (1998) the dollar value cost of fraudulent credit card transactions is incorporated into the model as a cost for making an incorrect prediction. A multi classifier meta learning approach is used to handle the non-uniform distribution of the samples. García, et al. (2007) provide an in-depth review of the important research in the area of class imbalance in pattern learning and classification.

3. CLASSIFICATION

The next task was to analyse the data selected for use in the development and training of a model and then to classify each sample as either bullying or not bullying. Cyberbullying can take many forms such as flaming, outing, exclusion, flooding, cyberstalking, impersonation or masquerading, trolling and denigration of their victims. These attacks could then be categorised as sexual, racial, cultural or against the intelligence or physical appearance of a person or their socio-economic status to name but a few. The types and categories of cyberbullying are further examined and analysed to develop a sense of the criteria against which to evaluate each sample. Once evaluated, each sample will then be classified as bullying or not bullying. The samples to be classified were scraped from the Ask.fm social networking website that uses a question and answer format to allow its users to interact.

3.1. Cyberbullying Criteria

This section provides more details about the type sample that could be considered as cyberbullying.

Flaming is usually considered as off topic insults or attacks against an individual or group and are more typically seen in a bulletin board or discussion type forums. A classic example of flaming is where a discussion comparing the virtues of the Linux, Microsoft Windows and Apple Mac operating systems disintegrates into a series of posts where each side questions the intelligence of the other two groups for using such an obviously inferior product. In the context of the Ask.fm dataset, it is unlikely that a series of posts of this type would be obviously detectable. The nature of the question and answer format, the order of the answers given, and their intermingling with other question would prevent it. However, even if a series of questions that are flaming in their nature, either against an individual or a number of individuals, cannot be specifically identified as flaming, their hostility, aggressiveness or derogatory sentiments would probably guarantee that they are identified as bullying. A user that purposefully starts,

encourages or engages in such flaming activities is sometimes known as a *Troll* and by denigrating or annoying people in this manner they are said to be *Trolling*.

The term ***outing*** is typically used to describe the disclosure of a person's sexual orientation, for example, revealing for the first time that a person is gay. As the users of the Ask.fm site are mostly teenagers and young adolescents, and because of the amount of sarcasm and black humour observed, it would be difficult reading isolated questions to determine whether a genuine outing has occurred. There is no facility to validate any such outing claim, or to determine if the question is just a slanderous attack. Regardless of these uncertainties, any post that appears to out someone or to question their sexual orientation must be considered as bullying.

Cyberexclusion can be considered as any type of activity that purposefully excludes an individual or group from an event, a party, shopping trip or holiday, for example. Alternatively, perhaps, they are prevented from joining a sports team or social group. This exclusion could take the form of inviting everyone to a party but then, explicitly, uninviting the excluded person in a public and humiliating manner. Another scenario could be to openly mock or taunt someone because they were prevented from joining a popular social clique in school. Any such obvious exclusionary activity seen in the Ask.fm dataset would be highlighted as bullying.

In some ways ***flooding*** could be seen as similar in nature to a Denial of Service (DoS) attack on a website. A DoS attack is an attempt to interrupt the normal functioning of a website or make it unavailable to users by bombarding the site with so many requests that the sites response time to legitimate traffic is extremely slow or non-existent, rendering the site unusable. In a chat room, flooding could be seen as repeatedly, and in very quick succession, posting empty messages or the same message again and again to prevent other users in the chat room from voicing their opinion. Alternatively, in the context of the Ask.fm dataset, flooding could be considered as the repeated asking of the same question or requests for information. Such flooding activity, if seen, should be considered bullying. However, as the data has been both anonymised and randomised, or instances of flooding may already have been deleted by the targeted user, it may be difficult to identify.

Stalking is seen as the unwanted, obsessive and criminal intrusion into the personal life of an individual by another person. When computers or other electronic equipment is used it can be considered as ***cyberstalking***. The actions of a stalker can include activities such as a desire to control their victim, to subversively gather information about them, monitoring and anonymously commenting on their on-line activity, making claims of false victimisation against the target and other false accusations such as defamation and libel. Other behaviours exhibited by the stalker may include flooding the victims with constant requests for personal information or for requests to meet in person. Any questions that are malicious in nature, from their content or tone, should be considered bullying. However, some questions, which may at first appear innocuous, for example "how old are you", "where were you last night" or "send me a selfie" should also be classified as bullying as they are requesting personal information which could be used by a potential stalker.

Impersonation or ***masquerading*** is where a person creates a false identity, or assumes the identity of someone else, when posting or creating on-line content. Typically, the reason for posting content using this assumed or false identity is to harm the reputation of the person they are pretending to be, by making what appear to be personally humiliating comments or by making harmful, harassing, confrontational or bullying remarks against another person. Alternatively, the owner of a false account could use it solely for the purpose of anonymously posting harmful content while observing on-line etiquette with their real name account. Due to the anonymous nature of the content on the Ask.fm site and the lack of any validation of a person's true identity it would be difficult if not impossible to ascertain whether a question is asked by a user masquerading or impersonating another person. However, it could also be argued that regardless of who posted the question if it meets any of the criteria that identify cyberbullying then it should be marked as such.

It is also important to consider the tone and content of a cyberbullying incident as well as the various forms that it may take. As well as threats of physical violence, the wishing of harm on a person, or by goading a person to harm themselves, the tone of a cyberbullying attack can be categorised in many ways including gender, sexist, racial, cultural, nationality, ethnicity, colour or race, intellectual, appearance, religious or socio-economic. Questions that could be considered confrontational, libellous, defamatory or make unfounded accusations that are intended to be hurtful or that could result in unstated repercussions should be considered as bullying.

Gender or sexist attacks would typically target women or sexual minorities, for example people of gay, lesbian, bisexual or transgender orientation. Simple examples of this type of cyberbullying would be name calling such as referring to a person as a *"bitch"*, *"faggot"* or *"dyke"* or questions relating to sexual experience. Other more sinister types of these attacks would include unwanted sexual advances or invitations to engage in or descriptions of unsolicited sexual activity. An extreme case would be the threat or implication of rape. Racial or cultural cyberbullying includes slurs or attacks against a cultural minority and its traditions, or direct attacks against a person based solely on their skin colour, ethnicity, race or nationality. Questions asking someone in a derogatory way if they are mentally or physically handicapped or questions that are hurtful or mean about their weight, height or general appearance are all cyberbullying.

Though difficult to identify in isolation, text that may be considered grooming in nature should also be identified. There are multiple different categories of grooming approaches including Approach, Communicative Desensitisation, Compliment, Isolation and Reframing. Samples of each type of categories are "i just want to meet", "how cum", "you are a really cute girl", "are you alone" and "let's have fun together" (Kontostathis, Edwards and Leatherman 2009).

It should be noted that profanity, without any of the other criteria listed here, would not, for the purposes of this research, be considered as cyberbullying.

The criteria for identifying cyberbullying can be summarised as follows:

- Can readily be identified as being attempts at flaming or of flooding.

- Is an obvious attempt to exclude or isolate an individual from a group or event.
- Could be perceived as stalking by overtly asking for personal information or photos.
- May be considered grooming where the question could be seen as an approach, an attempt to isolate or any of the other categories defined.
- Unnecessarily sexually explicit references or unsolicited invitations or by outing an individual whether real or perceived.
- Derogatory comments about race, culture, nationality, ethnicity or religion.

3.2. Dataset Classification

Once the criteria to be used to identify a question as bullying had been determined, the next step was to classify the block of questions that were to be used to train and test the initial model. A dataset of approximately 11,000 sample records, questions, were scrapped from the Ask.fm social networking site. Classification of this dataset was a two-step process. First, all questions were manually reviewed and classified as either bullying, not bullying or discard. In total, 10,914 question were reviewed. 1,644 were classified as bullying, 78 were classified as discard and the remainder as not bullying. Questions that were classified as discard contained only digits, emoticons or, what appeared to be, random characters.

To validate that the classification criteria specified are both meaningful and understandable and that the original classification was repeatable, accurate and consistent, a random selection of 500 questions, 250 bullying and 250 not bullying, were independently classified by two other reviewers. Each reviewer was given the classification criteria described above and a print out of the original unedited questions. They were then asked to classify each sample question as either bullying or not bullying.

In total, the first reviewer classified 246 questions as bullying of which 230 were originally identified as bullying. The second reviewer predicted 291 questions as bullying, of which 242 were originally identified as bullying. In total 224 of 250 sample questions, or just under 90%, were identified by all three classifiers as bullying. However, 248 out of 250 sample bullying questions, or 99.2%, were identified as bullying by two out of the three classifiers. Of note, 14 sample questions, that were originally classified as not bullying, were classified as bullying by both the independent reviewers. At just over 5.5% this is not hugely significant but important to highlight. Overall, however, this was a very satisfactory validation of both the classification criteria and the original classification of the dataset.

Before leaving this section, it is worth examining some of the questions that were classified as bullying to understand the rational of their classification. All the samples listed here are genuine examples from the Ask.fm dataset that was classified. Some of the samples have been abbreviated so as to highlight the relevant part of the question.

- Asking someone their age. Asking how old they are could be very innocent and a genuine friendly request. However, it could just as easily be considered as stalking or as grooming.

  - Age?
  - How old are you?
- Asking for information of an obviously private nature
  - What Bra size are you?
  - Can i have ur number?
  - How many ladies have you slept with?
- Asking someone to post a picture or video of themselves often requesting specific parts of the body, clothes or naked.
  - Puberty progress photo?
  - Bikini picture?
  - Pap of you right now?
- Asking someone about their sexual experience or preferences
  - Are you a virgin?
  - Are you good at giving head?
  - Butt or boobs
- Threatening physical violence or goading others to hurt themselves
  - Cut yourself ugly saggy ass hoe
  - I'm gonna beat your trashy little twig of an ass in so far you'll puke it out
  - I will slap you across your f**king mouths

It is important to note that these are only a small example of the questions that were classified as bullying. The content of a large proportion of the questions would be considered crude, vulgar and too offensive to list here.

4. Initial Data Modelling

The focus of this section is to describe the process to develop an initial model for predicting whether a question from the Ask.fm website could be considered as either bullying or not bullying. This model will then be used as a ground truth against which new models, generated to tackle the class imbalance issue, can be rated. Before describing the initial simple model developed it is worth providing a review of the original dataset and some transformations performed on the original dataset in order to explore multiple parallel lines of enquiry.

4.1. Data Description

Ask.fm is a social networking website that uses a question and answer format to allow its users to interact. Each post on a user's stream has a number of data attributes including the question

that was asked. A dataset of approximately 11,000 sample records, questions, were scrapped from the this site.

In total seven versions of the classified dataset were produced:

- **Dataset_01**
  This dataset is raw question data as extracted from Ask.fm that contains the original question text as scraped.
- **Dataset_02**
  A version of dataset_01 that has been cleaned. This dataset was considered the primary dataset. In total eight transformation steps were applied to the original question text:
  - Convert to lower case
  - Convert to ASCII
  - Remove URLs
  - Remove all punctuation
  - Replace numeric values with text values
  - Remove any digits that remain
  - Remove any unnecessary repeating characters to return valid words
  - Fix common abbreviations and replace with full word
- **Dataset_03**
  Bigrams created from the cleaned dataset [dataset_02]
- **Dataset_04**
  Trigrams created from the cleaned dataset [dataset_02]
- **Dataset_05**
  The cleaned dataset (dataset_02) with stop words removed
- **Dataset_06**
  Bigrams created from dataset_05
- **Dataset_07**
  Trigrams created from dataset_05

4.2. Model Evaluation

When evaluating the performance of a model in predicting whether a question was correctly classified as bullying or not, precision, recall and accuracy was used. Precision and recall are inversely related, meaning that as precision increases recall decreases and inversely where recall increases precision decreases. When developing a classification model, the critical decision is whether to seek to have high precision and low recall or to develop a model that delivers a low precision value but has high recall. Consider a scenario where we are trying to classify questions as bullying. High precision and low recall values suggest that a high percentage of questions predicted as bullying will be bullying. However, a significant number of bullying questions will not be correctly identified. A high recall value implies that a large percentage of bullying questions have been correctly identified but, as a consequence, a large number of not bullying questions would also be incorrectly identified as bullying yielding low

precision. Usually a trade-off has to be made between precision and recall depending on the situation and the preferred outcomes. G-Performance measurement was also calculated.

4.3. Initial Modelling

A Naïve Bayes learner, using a feature-based bag of words approach, was chosen as the learner to be used to develop an initial model against which the different approaches to the class imbalance could be evaluated. The Python based Natural Language Tool Kit (NLTK) was chosen to implement the learner. The model developed was then run on all seven datasets. The performance of the model for each dataset are shown in **¡Error! No se encuentra el origen de la referencia.**.

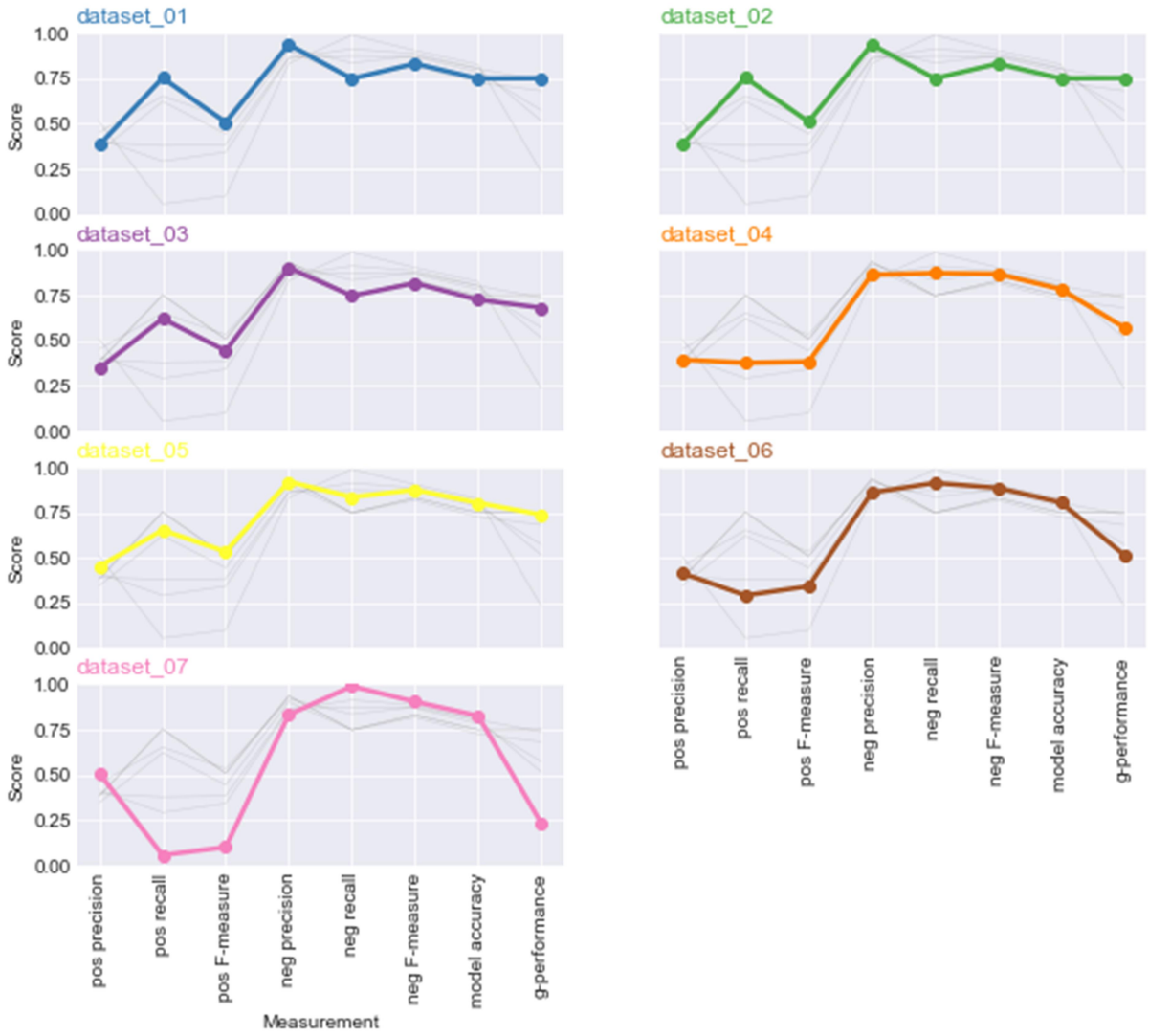

FIGURE 1. PERFORMANCE OF THE INITIAL NLTK MODELS.

The first observation from an analysis of the data is that the precision of the model in predicting the positive class, the bullying questions, is poor when compared to the values obtained for the negative class. The average precision for the positive bullying class is 41% compared to an average value of 89% for the not bullying class. Although the average performance for recall of

the positive class is 50%, this figure is not representative because of the value achieved for dataset_07, trigrams with stop words removed, which is less than 10%.

Taking all performance measures into account, it could be said that the best performance achieved by this model was with dataset_05, unigrams with stop words removed. However, with a precision value for the positive class of 45.3% and a recall value of 65.2%, the results are far from satisfactory. This was just an initial modelling attempt to establish a ground truth for the different class imbalance approaches. There are many changes that could be made to this model to improve its performance. Apart from addressing class imbalance, only selecting the features that have the highest impact and by changing from the feature-based approach offered by NLTK to the more advanced TF-IDF implementation offered by the Scikit-Learn package are some of the options that were considered.

5. Class Imbalance

It was observed that there was a class imbalance between the positive and negative samples in the datasets. In this section, it will be investigated if oversampling the minority positive class or undersampling the majority negative class affects the performance of the initial model. A compromise solution where the positive class is partially oversampled, and the negative class is partially undersampled, is also examined.

5.1. Majority Class Undersampling

Undersampling of the majority negative class was the first approach to be explored. In all datasets the ratio of negative, not bullying, to positive, bullying, classes is approximately 4.8:1. To determine how different negative class to positive class ratios affect the model performance, undersampling of the majority class at ratios of 3:1, 2:1 and 1:1 to the minority class were examined. Because the negative samples are randomly chosen to make the desired ratios, it was important to run multiple executions for each dataset to allow for any possible variance or imbalance in the samples selected. It was found that five executions were enough to show that all results were similar and consistent. The performance results achieved for each dataset at ratios of 3:1, 2:1, 1:1, and the original performance measures from the initial modeling, are shown FIGURE **2**.

It is clear that the performance of the positive class prediction is increasing as the ratio of samples approaches 1:1. The positive recall values for some of the datasets also show modest improvement, however, the values do not show the same improvement for dataset_04 trigrams, dataset_06 bigrams no stop words and dataset_07 trigrams no stop words. This lack of improvement could be attributed to the uniqueness of the bigrams and trigrams in the datasets. It was observed that the average frequency of n-grams in these datasets was very close to 1. Also, of note, is that as the ratio of classes approaches 1:1 the negative class precision and recall values are decreasing.

At a ratio of 1:1 the overall performance of dataset_03, bigrams including stop words, could, at this early stage of development, be considered very satisfactory. With performance values in all categories of just under 70% the model is equally accurate predicting both positive and negative

classes. If the ability to solely predict the positive class was the main driving force, then both unigrams models, with a sample ratio of 1:1, offer a better solution but at the cost of over predicting samples as positive. It must be kept in mind that one of the major drawbacks of undersampling in this manner is the risk of discarding samples that may, in fact, be very representative of the general population.

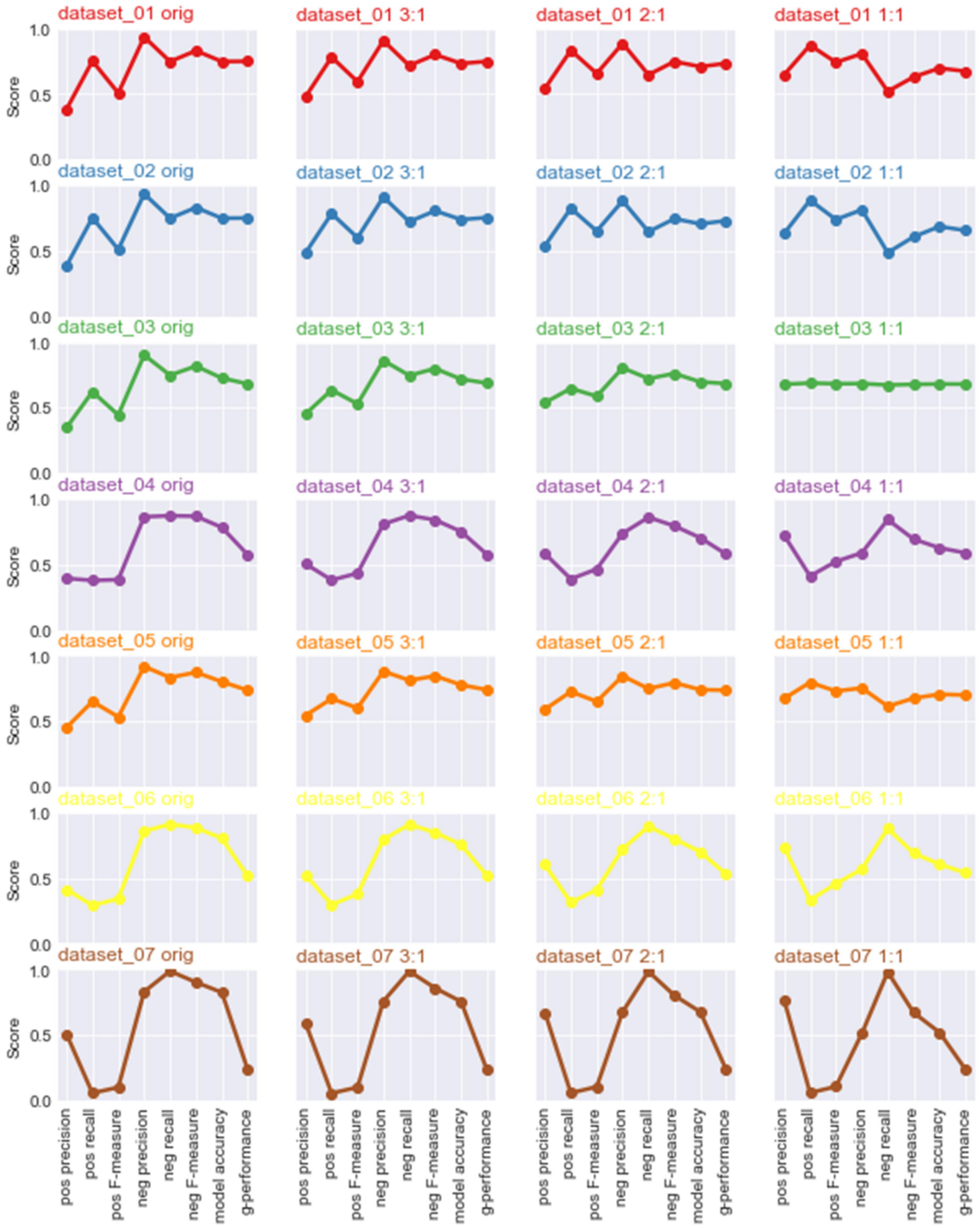

FIGURE 2. MODEL PERFORMANCE: UNDERSAMPLING OF MAJORITY CLASS.

5.2. Minority Class Oversampling

Oversampling of the positive minority class was explored next. Ratios of 3:1, 2:1 and 1:1 were again simulated, but this time, instead of reducing the number of negative samples in order to achieve these ratios the number of positive samples was increased. In line with all testing to this point the method used to increase the number of positive samples was as simple as possible.

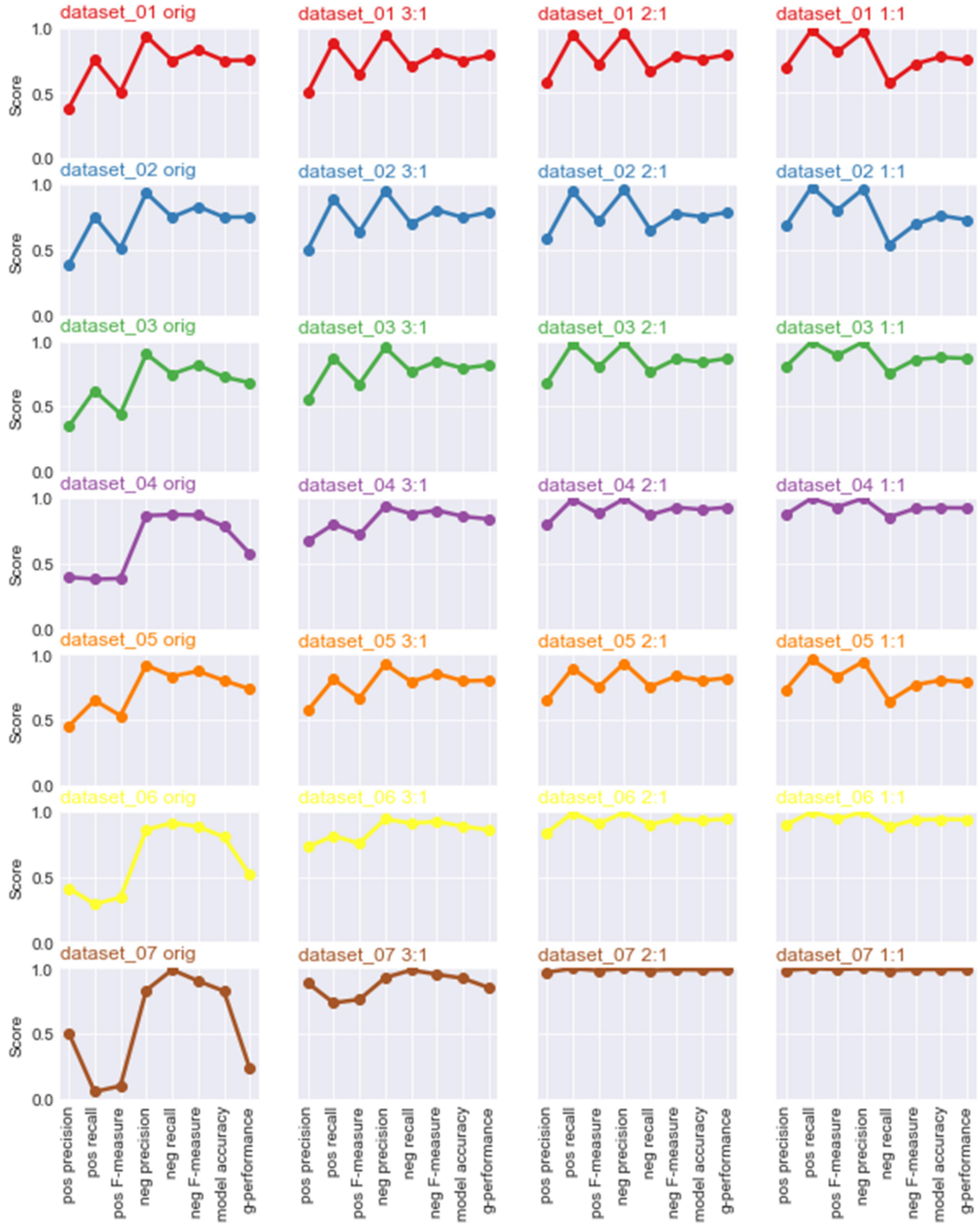

FIGURE 3. MODEL PERFORMANCE: OVERSAMPLING OF MINORITY CLASS.

When undersampling, the negative class was shuffled and the samples to be modeled chosen at random. This time the positive class was shuffled and the number of times the positive samples needed to be replicated, to achieve the desired ratio, is calculated. As before, the model was generated five times to ensure that an average performance was achieved. The performance results from the models generated using undersampling of the majority class with the models generated using oversampling are shown in FIGURE **3**.

It is clear that oversampling of the minority positive class has significantly improved the performance of all models. Bigram and tri-gram tokens, with stop words removed, nearly achieving perfection with 99.8% and 100% positive sample recall and 88.4% and 98.4% negative sample recall respectively. Inversely though, it was also observed that as the ratio of the classes approaches 1:1 that the recall performance of the model predicting the negative class decreased significantly, particularly for all the unigram datasets.

Overall though, the results achieved using oversampling of the minority class outperform the models developed using undersampling of the majority class. It must be questioned though whether this gain in performance has been achieved by overfitting the models to the repeated samples? This overfitting to replicated data is a known issue when oversampling is applied.

5.3. Hybrid Approach

The third option used to tackle the imbalance of the positive and negative classes was a hybrid approach that used both undersampling of the majority class and oversampling of the minority class. In this hybrid method, the normal approach is to take the total number of examples and then divide by the number of classes to get the target sample number. For example, in a two-class scenario with 800 positive and 400 negative samples, the target number of samples for each class would be 600 if a ratio of 50:50 was the desired ratio. In this section, both classes are over or undersampled to half the total number of examples as just described. However, as the negative class is significantly superior to the negative sample, ratios of 60:40 and 70:30 were also explored. The performance results of these hybrid models are shown in FIGURE **4**.

Comparing the performance of the model from oversampling of the minority class and the model from the hybrid approach, it is clear that the results achieved by oversampling the minority class to ratios of 1:1 and 2:1 are very similar to the results achieved by the 50:50 and 60:40 hybrid approach. This fact is highlighted by TABLE **1**. The top part of this table gives the results for the oversampling model where a ratio of 1:1 was used. The middle part of the table gives the performance results for the model where hybrid sampling with a ratio of 50:50 was used. The bottom part of the table gives the percentage difference between the oversampling model and the hybrid sampling model. It can be seen that the majority of all performance measures between the two models are within 1% of each other. Considering all performance measures, the total average performance difference between all the models of each type is just 0.18%. Although the difference between the 2:1 oversampling model and the 60:40 hybrid model is slightly more obvious again the total average difference between the two models is just 1.78%.

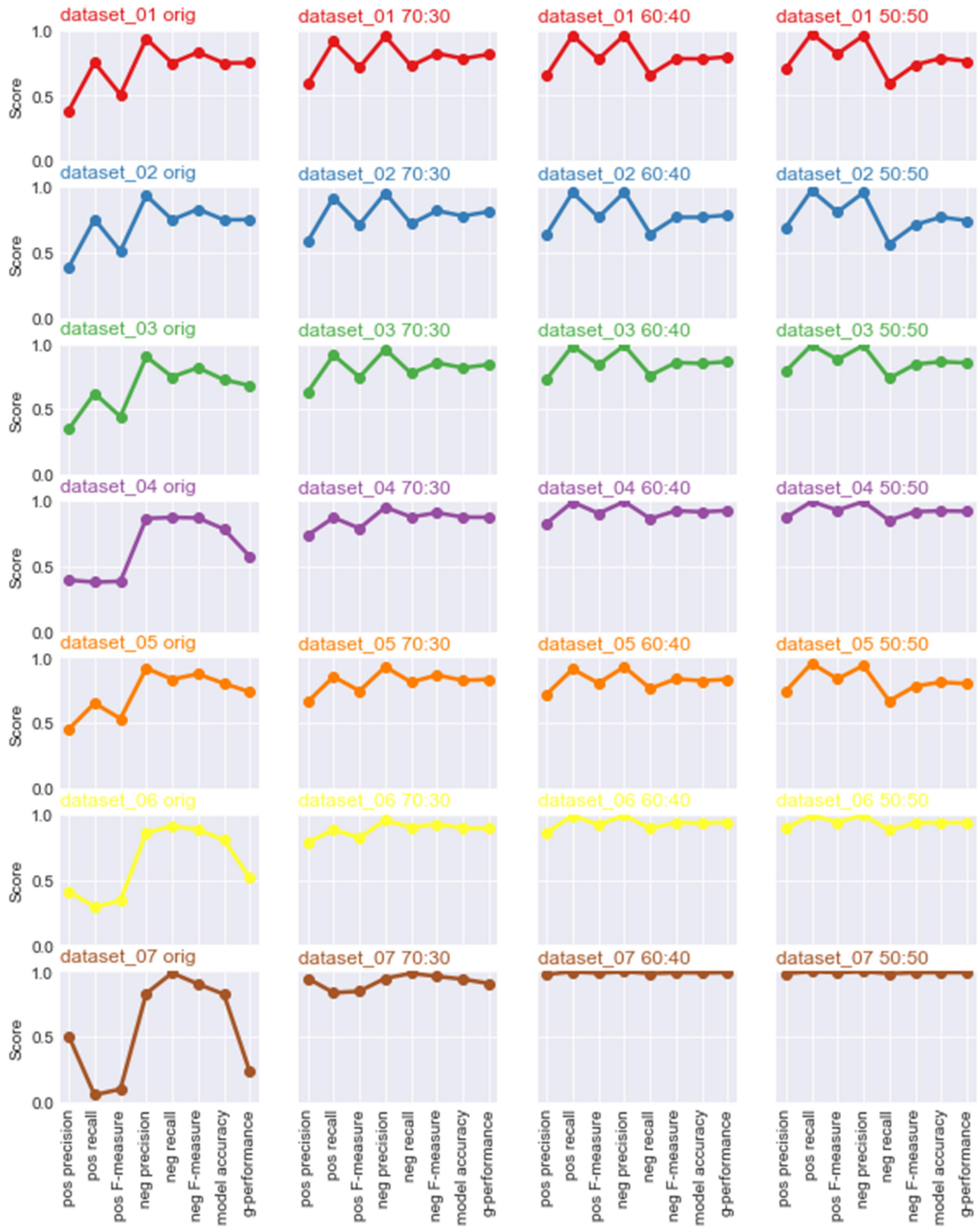

FIGURE 4. MODEL PERFORMANCE FOR HYBRID APPROACH.

TABLE 1. OVERSAMPLING AND HYBRID SAMPLING PERFORMANCE COMPARISON.

Oversampling Minority Class – Ratio 1:1							
Measurement	dataset 01	dataset 02	dataset 03	dataset 04	dataset 05	dataset 06	dataset 07
pos precision	0.698	0.682	0.804	0.872	0.733	0.896	0.984
pos recall	0.981	0.981	0.997	0.998	0.965	0.998	1.000
pos F-measure	0.815	0.804	0.890	0.930	0.833	0.944	0.992
neg precision	0.968	0.966	0.996	0.998	0.949	0.997	1.000
neg recall	0.574	0.542	0.756	0.853	0.647	0.884	0.984
neg F-measure	0.720	0.693	0.859	0.919	0.769	0.937	0.992
model accuracy	0.777	0.761	0.876	0.925	0.806	0.941	0.992
g-performance	0.750	0.729	0.868	0.923	0.790	0.939	0.992
Hybrid Sampling – Ratio 50:50							
Measurement	dataset 01	dataset 02	dataset 03	dataset 04	dataset 05	dataset 06	dataset 07
pos precision	0.707	0.692	0.791	0.870	0.744	0.894	0.984
pos recall	0.975	0.975	0.992	0.995	0.957	0.996	1.000
pos F-measure	0.819	0.809	0.880	0.928	0.837	0.942	0.992
neg precision	0.959	0.957	0.989	0.994	0.940	0.996	1.000
neg recall	0.595	0.566	0.737	0.850	0.670	0.881	0.984
neg F-measure	0.734	0.711	0.844	0.917	0.782	0.935	0.992
model accuracy	0.785	0.770	0.864	0.923	0.813	0.939	0.992
g-performance	0.762	0.743	0.855	0.920	0.801	0.937	0.992
Percentage Performance Difference							
Measurement	dataset 01	dataset 02	dataset 03	dataset 04	dataset 05	dataset 06	dataset 07
pos precision	1.28%	1.49%	-1.64%	-0.23%	1.51%	-0.24%	-0.02%
pos recall	-0.64%	-0.64%	-0.51%	-0.30%	-0.83%	-0.13%	-0.03%
pos F-measure	0.49%	0.61%	-1.14%	-0.26%	0.49%	-0.19%	-0.02%
neg precision	-0.89%	-0.90%	-0.69%	-0.34%	-0.93%	-0.15%	-0.03%
neg recall	3.71%	4.47%	-2.55%	-0.26%	3.51%	-0.31%	-0.02%
neg F-measure	2.04%	2.53%	-1.76%	-0.30%	1.69%	-0.24%	-0.02%
model accuracy	0.96%	1.18%	-1.39%	-0.28%	0.91%	-0.21%	-0.02%
g-performance	1.51%	1.89%	-1.53%	-0.28%	1.32%	-0.22%	-0.02%

6. Conclusions

The contributions of this paper are twofold. First a new dataset for use in cyberbullying research and the criteria used in its classification is presented. Next it was shown that, when presented with an imbalanced dataset, improved model performance can be achieved using undersampling of the majority class, oversampling of the minority class or a hybrid approach that uses a combination of both.

Following a review of the literature, an overview of the process used in the classification of each question as being either bullying or not bullying was given. Also included were detailed descriptions of the various types and categories of cyberbullying. Samples of some of the questions classified as bullying were also shown. Just under 11,000 sample questions were classified. The limited availability of this type of dataset for research was also noted.

When the dataset was examined, it was observed that there was a class imbalance between the positive and negative samples. It was investigated if oversampling the minority positive class or undersampling the majority negative class affected the performance of the model. In each case datasets in the ratios of 3:1, 2:1 and 1:1, majority to minority class were created. A compromise solution where the positive class is partially oversampled, and the negative class is partially undersampled was also examined. Datasets with ratios of 70:30, 60:40 and 50:50, majority to minority, were created for this approach. To add additional context multiple dataset with and without stop words and containing unigrams, bigrams and trigrams were examined.

Undersampling of the majority class showed the least improvement in model performances when compared to the original, unaltered, dataset. With an overall performance accuracy of just under 70% the best combination when undersampling the majority class was with a ration of 1:1, bigrams and including stop words. When using undersampling in this manner the risk of discarding samples that would be very predictive in the general population was shared.

The best results were given using oversampling of and majority class and the hybrid approach. Oversampling of the minority positive class showed a significantly improved model performance, especially where the dataset contained bigrams or trigrams and had stop words removed. It must be questioned though whether this gain in performance has been achieved by overfitting to the repeated samples? This overfitting to replicated data is a known issue when oversampling is applied. Comparing the performance of the models from oversampling of the minority class and the models from the hybrid approach, it is clear that the results achieved by oversampling the minority class to ratios of 1:1 and 2:1 are very similar to the results achieved by the 50:50 and 60:40 hybrid approach.

References

Cardie, Claire. 1997. "Improving minority class prediction using case-specific feature weights." *Proceedings of the Fourteenth International Conference on Machine Learning.* Morgan Kaufmann. 57-65.

Chan, Philip K., and Salvatore J. Stolfo. 1998. "Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection." *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining.* AAAI Press. 164-168.

Chawla, Nitesh V. and Bowyer, Kevin W. and Hall, Lawrence O. and Kegelmeyer, W. Philip. 2002. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research.* 321-357.

Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety." *Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).* IEEE. 71-80.

Cionnaith, Fiachra Ó. 2012. *Third suicide in weeks linked to cyberbullying.* Accessed 03 14, 2019. http://www.irishexaminer.com/ireland/third-suicide-in-weeks-linked-to-cyberbullying-212271.html.

Dadvar, M. , F. M. G. de Jong, R. J. F. Ordelman, and R. B. Trieschnigg. 2012. "Improved cyberbullying detection using gender information."

Dadvar, Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. "Improving Cyberbullying Detection with User Context." In *Lecture Notes in Computer Science*, 693-696. Springer Berlin Heidelberg.

Dadvar, Maral, Roeland Ordelman, Franciska de Jong, and Dolf Trieschnigg. 2012. "Towards User Modelling in the Combat against Cyberbullying." *Lecture Notes in Computer Science*, 277-283.

Dinakar, Karthik, Roi Reichart, and Henry Lieberman. 2011. "Modeling the Detection of Textual Cyberbullying." *The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011.* Association for the Advancement of Artificial Intelligence.

FBM, Fundación Barcelona Media. 2009. *CAW 2.0 Training Datasets.* Barcelona.

García, Vicente, José Sánchez, Mollineda R.A, Roberto Alejo, and José Sotoca. 2007. "The class imbalance problem in pattern classification and learning." *II Congreso Español de Informática.*

Kontostathis, April, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. "Detecting Cyberbullying: Query Terms and Techniques." *Proceedings of the 5th Annual ACM Web Science Conference.* New York: ACM. 195-204.

Kontostathis, April, Lynne Edwards, and Amanda Leatherman. 2009. "ChatCoder: Toward the Tracking and Categorization of Internet Predators." *Proc. Text Mining Workshop 2009 Held In Conjunction With The Ninth Siam International Conference On Data Mining (Sdm 2009). Sparks, Nv. May 2009.*

Kubat, Miroslav, and Stan Matwin. 1997. "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection." *Proceedings of the Fourteenth International Conference on Machine Learning.* Morgan Kaufmann. 179-186.

Nahar, Vinita, Xue Li, and Chaoyi Pang. 2013. "A step towards combating cyberbullying: Automated detection."

Nahar, Vinita, Xue Li, and Chaoyi Pang. 2013. "An Effective Approach for Cyberbullying Detection." *Communications in Information Science and Management Engineering.* 238-247.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers Inc.

Reynolds, K., A. Kontostathis, and L. Edwards. 2011. "Using Machine Learning to Detect Cyberbullying." *2011 10th International Conference on Machine Learning and Applications and Workshops (ICMLA).* Honolulu. 241-244.

Riegel, Ralph. 2013. *Cyber-bullies claimed lives of five teens.* 25 01. Accessed 03 14, 2019. http://www.herald.ie/news/cyberbullies-claimed-lives-of-five-teens-29043544.html.

RuleQuest Research. n.d. *Data Mining Tools See5 and C5.0.* Accessed 03 2013. https://www.rulequest.com/see5-info.html.

Smith-Spark, Laura. 2013. *Hanna Smith suicide fuels calls for action on Ask.fm cyberbullying.* 09 08. Accessed 03 14, 2019. http://www.cnn.com/2013/08/07/world/europe/uk-social-media-bullying/index.html.

U.S. Department of Health and Human Services. 2018. *What Is Bullying.* 26 06. Accessed 03 31, 2019. https://www.stopbullying.gov/what-is-bullying/index.html.

Weiss, Gary, Kate McCarthy, and Bibi Zabar. 2007. "Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?" *Proceedings of the 2007 International Conference on Data Mining, DMIN 2007.* Las Vegas: CSREA Press. 35-41.

Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. "Learning from Bullying Traces in Social Media." *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Stroudsburg: Association for Computational Linguistics. 656-666.

Xu, Jun-Ming, Xiaojin Zhu, and Amy Bellmore. 2012. "Fast Learning for Sentiment Analysis on Bullying." *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining.* Beijing: ACM. 10:1-10:6.

Yin, Dawei, Brian Davison, Zhenzhen Xue, Liangjie Hong, April Kontostathis, and Lynne Edwards. 2009. "Detection of Harassment on Web 2.0." *Proceedings of the Content Analysis in the WEB.* 1-7.