

ABSTRACT

This thesis explores the classification of natural disaster tweets, a topic that has been extensively researched in the past. However, previous studies did not address a key challenge in supervised learning: the need for large labeling datasets, which are time-consuming and laborious to create. To address this issue, we propose a solution using a weak supervision method. While weak supervision has been explored in previous research, it has not been applied specifically to the context of natural disasters. We believe that weak supervision can be particularly useful for training datasets related to understanding and emergency response in time-critical situations.

In our research, we adopt an Incomplete Supervision approach, leveraging a small initial dataset to construct a larger training dataset. We utilize a Labeling Function Generator with a query-based programmatic function to build the training dataset. To evaluate the effectiveness of our weak supervision method, we compare it with supervised learning using LSTM and SVM as classifiers.

Our findings indicate that, although supervised learning outperforms weak supervision in terms of performance, achieving a 93% figure compared to weak supervision's 90% figure in SVM, there are other advantages to consider. Specifically, our weak supervision method allows us to label a significantly larger number of datasets, approximately 26,000 compared to the 6,000 datasets labeled through supervised learning. Despite the performance gap, these advantages make the weak supervision approach a viable alternative.

Keyword – *NLP, weak supervision, natural disaster, text classification, twitter, SVM, LSTM*

INTISARI

Penelitian ini menjelajahi klasifikasi cuitan bencana alam, sebuah topik yang telah banyak diteliti sebelumnya. Namun, penelitian sebelumnya tidak mengatasi tantangan utama dalam pembelajaran berbimbing: kebutuhan akan dataset labeling besar, yang memakan waktu dan tenaga untuk dibuat. Untuk mengatasi masalah ini, kami mengusulkan sebuah solusi menggunakan metode supervisi lemah. Meskipun supervisi lemah telah dijelajahi dalam penelitian sebelumnya, namun belum diterapkan khusus untuk konteks bencana alam. Kami percaya bahwa supervisi lemah dapat sangat berguna untuk melatih dataset terkait pemahaman dan tanggapan darurat dalam situasi yang mengharuskan waktu cepat.

Dalam penelitian kami, kami mengadopsi pendekatan Supervisi Tidak Lengkap, memanfaatkan dataset awal yang kecil untuk membangun dataset pelatihan yang lebih besar. Kami menggunakan Generator Fungsi Label dengan fungsi program berbasis kueri untuk membangun dataset pelatihan tersebut. Untuk mengevaluasi efektivitas metode supervisi lemah kami, kami membandingkannya dengan pembelajaran berbimbing menggunakan LSTM dan SVM sebagai pengklasifikasi.

Temuan kami menunjukkan bahwa, meskipun pembelajaran berbimbing unggul dalam hal performa, mencapai angka 93% dibandingkan angka 90% dari supervisi lemah pada SVM, terdapat keuntungan lain yang harus dipertimbangkan. Secara khusus, metode supervisi lemah kami memungkinkan kami untuk melabeli jumlah dataset yang jauh lebih besar, sekitar 26.000 dibandingkan dengan 6.000 dataset yang diberi label melalui pembelajaran berbimbing. Meskipun terdapat perbedaan dalam performa, keuntungan-keuntungan ini membuat pendekatan supervisi lemah menjadi alternatif yang layak dipertimbangkan.

Keyword – *NLP, weak supervision, natural disaster, text classification, twitter, SVM, LSTM*