Streszczenie

Z powodu szybkiego wzrostu liczby artykułów naukowych publikowanych co roku coraz trudniej jest nadążać za rozwojem nawet tylko swojej dziedziny nauki. Badacze, a także np. urzędnicy decydujący o przydziale środków na badania naukowe polegają na tradycyjnych indeksach scjentometrycznych w celu wyszukiwania obiecujących lub potencjalnie przełomowych projektów badawczych. To podejście jest jednak obarczone pewnymi wadami.

Celem niniejszego projektu badawczego jest poszukiwanie rozwiązań niwelujących te wady i zaproponowanie automatycznej metody pomiaru innowacyjności publikacji naukowych poprzez predykcję ich wieku na podstawie analizy tekstu. Na niniejszą pracę składają się trzy recenzowane artykuły opublikowane w znaczących międzynarodowych źródłach opisujące postęp prac nad predykcją dat publikacji przy użyciu modelowania tematycznego, nadzorowanych modeli predykcyjnych i wreszcie aktualnych modeli osadzania słów (BERT) trenowanych na diachronicznych korpusach artykułów naukowych obejmujących wieloletnie okresy. Na bazie tych predykcji zaproponowano liczbową miarę odzwierciedlającą podobieństwo zawartości ocenianych artykułów do zawartości artykułów publikowanych w przyszłości lub przeszłości, a zatem ich prawdopodobną innowacyjność.

Proponowaną metodę zastosowano na trzech korpusach obejmujących publikacje ze źródeł wiodących w swoich dziedzinach. Pokazano, jak wartości proponowanej miary innowacyjności korelują z liczbą cytowań. Pokazano też na przykładzie dwóch korpusów obejmujących ponaddwudziestoletni okres, jak przy użyciu modeli BERT obniżyć średni błąd bezwzględny dla predykcji wieku publikacji odpowiednio z 3,56 i 2,56 roku do 0,68 i 0,64 roku.